Estimating errors in quantities of interest in the case of hyperelastic membrane deformation

Eleni Argyridou

Department of Mathematics, Brunel University London

February 6, 2018

ABSTRACT

There are many mathematical and engineering methods, problems and experiments which make use of the finite element method. For any given use of the finite element method we get an approximate solution and we usually wish to have some indication of the accuracy in the approximation. In the case when the calculation is done to estimate a quantity of interest the indication of the accuracy is concerned with estimating the difference between the unknown exact value and the finite element approximation. With a means of estimating the error, this can sometimes be used to determine how to improve the accuracy by repeating the computation with a finer mesh. A large part of this thesis is concerned with a set-up of this type with the physical problem described in a weak form and with the error in the estimate of the quantity of interest given in terms of a function which solves a related dual problem. We consider this in the case of modelling the large deformation of thin incompressible isotropic hyperelastic sheets under pressure loading. We assume throughout that the thin sheet can be modelled as a membrane, which gives us a two dimensional description of a three dimensional deformation and this simplifies further to a one space dimensional description in the axisymmetric case when we use cylindrical polar coordinates. In the general case we consider the deformation under quasi-static conditions and in the axisymmetric case we consider both quasi-static conditions and dynamic conditions, which involves the full equations of motion, which gives three different problems. In all the three problems we describe how to get the finite element solution, we describe associated dual problems, we describe how to solve these dual problems and we consider using the dual solutions in error estimation. There is hence a common framework. The details however vary considerably and much of the thesis is in describing each case.

This thesis is dedicated to my parents, Ivi and Lefkos, and my fiancé Petros.

ACKNOWLEDGEMENTS

First my thanks go to my principal supervisor Dr.Michael Warby for his guidance and support for both theoretical and numerical results and also for helping me in the preparation of the thesis. Secondly, I would like to thank my second supervisor Prof. John Whiteman for his assistance, encouragement and support during my research.

My thanks also go to Christopher Knapp, who was a great friend, was always willing to help and give his best suggestions.

Finally I owe a dept of gratitude to my parents, Ivi and Lefkos, and my fiancé Petros, who were supporting me and encouraging me during this period. Without their constant support and consideration, this accomplishment wouldn't have been made possible.

CONTENTS

1. INTRODUCTION

The finite element method (f.e.m) is one of the most powerful and widely used numerical methods for finding approximate solutions to mathematical problems formulated so as to simulate the responses of physical systems to various forms of excitation. It is used in various branches of engineering and science, such as elasticity, heat transfer, fluid dynamics, electromagnetics, acoustics, biomechanics etc.

In the f.e.m the solution domain is subdivided into elements of simple geometrical shape, such as triangles, squares, tetrahedra, hexahedra where a set of basis functions are constructed such that each basis function is non-zero over a small number of elements only. This is called discretization. The most popular method based on the discretization process is the Galerkin method which we introduce in chapter 2 and which we use throughout the thesis. In chapter 2 we just give a brief description of the finite element method. More analysis and description of the f.e.m can be found in the books of [5], [30], [9] and many others.

In this thesis, we focus on the formulation and solution of the discrete equations for nonlinear problems that are of principal interest in applications of the f.e.m to solid mechanics and structural mechanics. A typical approach of the nonlinear analysis as it is given in [9] can be described as follows.

1 Development of the model;

2 Formulation of the governing equations;

3 Discretization of the equations;

4 Solution of the equations;

5 Interpretation of the results.

The computational modelling of many engineering problems in solid mechanics involves the approximate finite element solution of the displacement field and, possibly also the

velocity field. Then, by using these f.e. approximations we are able to estimate some engineering quantities of interest (QoI). In this thesis, we consider the case of a thin sheet under pressure loading, which we model as a membrane. The QoI can be, for example, the localized stress in part of the sheet, the average thickness over a region, the potential energy of the deformed structure or the kinetic energy associated with the motion. To get to the stage when we have computed something sufficiently accurately we need a way of estimating the error in any given solution and a refinement procedure to attempt to determine how to repeat the procedure with a finer mesh in order to reach some desired level of accuracy. The details of what is done depends on whether we are most interested in the error in the primary function that we are approximating (e.g. the displacement field) or of some quantity of interest which is the goal of the computation. It is goal orientated techniques in the case of a deformation of a membrane under pressure loading which is the main topic of this thesis. More analysis of the error estimation techniques that can be used and adaptive mesh refinement can be found in the books of [5] and [1]. The goal oriented technique that is used in this thesis is based on the methods produced by Rannacher and his co-workers, see e.g. [6] and papers by Oden and his co-workers as in [19], [20], [17] and [18].

In the thesis we describe the physical problems being considered which is that of the inflation of a membrane under pressure loading. This is done in the general non-axisymmetric case under quasi-static conditions, the axisymmetric version also under quasi-static conditions and the axisymmetric case again but with the full equations of motion which we refer to as the dynamic case. For each of these three cases we describe the problem in a weak form and in each case we apply a technique, which we briefly describe shortly, which leads to a related dual problem which is the key to the goal orientated approach. There are different orders in which the material can be presented as, for example, whether all the equations related to the membrane model are given first and all the dual problem material is given later on or whether there is a different self-contained chapter for each of the three problems. The order that is chosen here is closer to the second of these possibilities although there is a separate chapter just on the creation of a dual problem in an abstract way in order to be able to represent the error in the approximation of a QoI. This is done so that the overall technique is clear before any specific case is described. With the membrane model already described at this stage the chapter then ends with the first of the three cases that we consider. The details of applying the technique to each cases is highly problem dependent which is why these parts are mostly separated. The contributions of the thesis is in showing how the technique can be applied and the results obtained for situations not considered in related work in [28]. There are hurdles to overcome in the non-axisymmetric case but otherwise

things worked out close to what might have been predicted. The dynamic case proved more difficult than first envisaged and the use of a high order scheme in time is needed when high accuracy is required.

The thesis is organized as follows. First, in chapter 2, we start with a brief description of the f.e.m, where we introduce many of the basic terms and we give preliminary material which is used later. There is material on one-dimensional basis functions and Legendre polynomials which are needed in the axisymmetric problems in the quasi-static and dynamic cases and there are details for two-dimensional problems including how to deal with the refinement of triangular meshes.

The description of the physical problem starts in chapter 3. In chapter 3, we move to the description of the f.e.m for a nonlinear problem, which involves a vector of unknowns. We consider the use of a membrane model of a thin sheet under pressure loading. For computational purposes, the membrane theory gives us a 2-dimensional description of a 3-dimensional deformation. In addition we assume that the membrane is composed of a homogeneous, isotropic, incompressible hyperelastic material. Our aim here is to first find a f.e. approximation of the displacement field ($\underline{u}$) of the membrane deforming under quasi-static conditions when a pressure is applied and from this to compute a QoI. The actual nonlinear problem that we wish to solve can be described as follows.
Find $\underline{u} \in V$ such that

$$A(\underline{u}; \underline{\alpha}) = 0 \quad \forall \underline{\alpha} \in V, \tag{1.0.1}$$

with V being the Hilbert space. It is written in this way with 0 on the right hand side as the pressure loading term depends on the displacement field $\underline{u}$ that we are trying to find and, as we show, the expression for $A(\underline{u}; \underline{\alpha})$ contains two parts which are respectively at term corresponding to the stress in the membrane and a terms corresponding to the pressure loading. The Hilbert space V is a subspace of $H^1(\Omega)$ where Ω is the undeformed membrane mid-surface or to be a bit more precise each component of $\underline{u}$ is in a subspace of $H^1(\Omega)$. This does fit in with the framework of what is done in of chapter 4 when we allow for a non-zero right hand side vector. We assume that (1.0.1) admits a unique solution $\underline{u} \in V$. The detail in chapter 3 is in setting up the membrane problem in (1.0.1), some detail relates to how to obtain the approximate solution by the finite element method.

Chapter 4 is where, in an abstract setting, we describe how to represent the error in a QoI involving a function which solves a related dual problem and we get an estimate of an error by approximating the exact dual problem. To be a bit more specific, let $\underline{u} \in V$ satisfy

$$A(\underline{u}; \underline{\psi}) = F(\underline{\psi}), \quad \forall \underline{\psi} \in V, \tag{1.0.2}$$

where V is an infinite dimensional function space. $A(\cdot;\cdot)$ represents a semi-linear form which is such that it is linear in arguments to the right of the semicolon and nonlinear in arguments to the left of the semicolon and $F(\cdot)$ represents a linear functional. Also let $J(\underline{u})$ denote the QoI we wish to compute with $J(\cdot)$ being a functional. If we obtain an approximation $\underline{u}_h$ to $\underline{u}$ from a finite element space then we can represent the error in the form

$$J(\underline{u}) - J(\underline{u}_h) = F(\underline{z}) - A(\underline{u}_h;\underline{z}) \tag{1.0.3}$$

where $\underline{z}$ is from an infinite dimensional space and satisfies

$$\int_0^1 A'(\underline{u}_h + s\underline{e}_h;\underline{\alpha},\underline{z})\,\mathrm{d}s = \int_0^1 J'(\underline{u}_h + s\underline{e}_h;\underline{\alpha})\,\mathrm{d}s \quad \forall \underline{\alpha} \in V, \tag{1.0.4}$$

where $A'(\cdot;\cdot,\cdot)$ and $J'(\cdot;\cdot)$ denote Gâteaux derivatives and where $\underline{e}_h = \underline{u} - \underline{u}_h$. We cannot obtain $\underline{z}$ as it is from an infinite dimensional space and the problem involves the unknown exact error $\underline{e}_h$. However we can consider approximating the problem given in (1.0.4) and the simplest approximation is to consider finding $\underline{z}_h \in \hat{V}_h$ such that

$$A'(\underline{u}_h;\underline{\alpha},\underline{z}_h) = J'(\underline{u}_h;\underline{\alpha}) \quad \forall \underline{\alpha} \in \hat{V}_h \tag{1.0.5}$$

for a suitable space $\hat{V}_h$ from which we get the estimate

$$J(\underline{u}) - J(\underline{u}_h) \approx F(\underline{z}_h) - A(\underline{u}_h;\underline{z}_h). \tag{1.0.6}$$

This is described further in chapter 4 with details in specific cases described in later chapters. References for this material can be found in the work of [22]. The general framework for nonlinear problems was advanced by [8]. See also [28] which contains work that we extend in this thesis. With the abstract setting given the chapter ends with the details with one if the three cases that is considered in this thesis. In particular, in the case of the problem described in chapter 3 there is detail relating to dual problems in that we give details of $A'(\cdot;\cdot,\cdot)$ and $J'(\cdot;\cdot)$ for various QoI. In the dual problem set-up we get the estimate

$$J(\underline{u}) - J(\underline{u}_h) \approx -A(\underline{u}_h;\underline{z}_h) \quad \forall \underline{z}_h \in \hat{V}_h \tag{1.0.7}$$

and we then investigate if this helps in which elements to refine in an adaptive refinement procedure. Most of the results in such tests are given in chapter 5.

In chapter 6, we give the description of the simplified nonlinear pressure model in the case of a hyperelastic axisymmetric circular disk. By using cylindrical polars, we describe the membrane model which is now reduced to one dimension and where the unknowns depend only on the radial dimension. We first consider the quasi-static case where, per-

haps not too surprisingly, it is comparatively easy to get high accuracy compared with the other cases presented in the thesis. Extending to the dynamic case is more involved as we have a description which has an expression for $A(\cdot;\cdot)$ which involves integrating in both space and time and the finite element scheme gives us an approximation $\underline{u}_h$ and $\underline{v}_h$ to respectively the displacement $\underline{u}$ and the velocity $\underline{v}$. In the scheme the velocity $\underline{v}_h$ and time derivative of $\underline{u}_h$ only match in a weak sense. The detail in chapter 6 is in giving the expressions involved so that the exact solution can be described as follows.
Find $\begin{pmatrix} \underline{u} \\ \underline{v} \end{pmatrix} \in V$ such that

$$A\left(\begin{pmatrix} \underline{u} \\ \underline{v} \end{pmatrix}; \begin{pmatrix} \underline{\psi} \\ \underline{\theta} \end{pmatrix}\right) = F\left(\begin{pmatrix} \underline{\psi} \\ \underline{\theta} \end{pmatrix}\right) \tag{1.0.8}$$

where $A(\cdot;\cdot)$ is a semilinear form on the Hilbert space V and $F(\cdot)$ is a linear functional on V. It is described in this way with the "test vector" in this context given in two parts in the expressions and we describe these as $\underline{\psi}$ and $\underline{\theta}$.

Next in chapter 7 we move to the goal-oriented technique which involves dual problems associated with the problems described in chapter 6 and in particular the detail involves the expression for the term $A'(\cdot;\cdot,\cdot)$. As we show we get a problem which is backward in time. Much of the chapter is concerned with the details that this involves. In [28] the dependence of $\underline{u}_h(r,t)$ and $\underline{v}_h(r,t)$ on time t is a degree 1 polynomial in t on an interval $t_{j-1} \leq t \leq t_j$ and we describe this case here and we can also do this in the corresponding dual problem. However, for the QoI considered much of the error is due to the time discretization and a very large number of time steps are needed which makes it quite expensive to get high accuracy. To improve the situation a higher order in time scheme is described for the approximations $\underline{u}_h$ and $\underline{v}_h$ and also for the unknowns in the related dual problem. Most of the results of these studies are given in chapter 8. An outcome of the work is that we can get high accuracy and the estimate of the error via the solution of a dual problem is a good estimate when the approximations to $\underline{u}$ and $\underline{v}$ are sufficiently good although the entire computation can be quite expensive.

2. BACKGROUND FINITE ELEMENT MATERIAL

2.1 Introduction

This chapter contains a collection of topics related to the finite element method which are needed later in the thesis. A reader already familiar with the topics should be able to skip all or much of the material and just refer back to specific things when the finite element method is being described for the three different cases of the membrane inflation problem considered in later chapters.

We start the detail with the idea of describing a problem in a weak form which is needed throughout the thesis and with the Galerkin method which is used to get an approximate solution. This is first done in an abstract way and then specific detail is given for one-dimensional and two-dimensional problems considered later.

In the preliminaries for the one-dimensional problems much of the detail is about the basis functions with details about Legendre polynomials which are used to describe basis functions in both space and time at various places in the thesis.

There is more preliminary material related to the two dimensional case with specific detail involving Poisson's problem and piecewise linears and piecewise quadratics on a triangular mesh. Piecewise linears and quadratics are both used when we later approximately solve the membrane inflation problems. We include known *a-priori* error estimates in order to describe how fast a sequence of finite element solutions converge to the exact solution using various norms as a mesh is refined. To improve a given approximation we refine the mesh and to do this in an efficient way we need appropriate *a-posteriori* error indicators to drive an adaptive refinement technique. *A-posteriori* error estimation in terms of a norm of the error in the primary unknown is not exactly the topic of this thesis but it is close to the goal orientated techniques that we consider in later chapters. Hence the detail of such techniques is kept brief. Given any estimates of how the error in something varies throughout a domain we need to cope with the detail of generating

a finer mesh from the coarser mesh and towards of the chapter we describe how this can be done and present some numerical results for a model problem to show how this works in practice.

2.2 Weak Formulation

2.2.1 General case of the weak form

First we give the general case of the weak form of a linear elliptic differential equation, given by a bilinear form $a(\cdot,\cdot)$ in a Hilbert space V.

Let $\Omega \subset \mathbb{R}^n$ for $n = 1, 2$ be a simply connected open set and let V be the Hilbert space of real valued functions defined on Ω.

Let $f : \Omega \to \mathbb{R}$ be a sufficiently continuously differentiable function such that we can consider the following differentiable equation

$$Lu = f \quad \text{in } \Omega, \tag{2.2.1}$$

with L being a linear elliptic differentiable operator. To able to uniquely solve we also have boundary conditions on u on the boundary $\partial\Omega$ of Ω.

Now, by multiplying the equation (2.2.1) with a suitable smooth test function v and integrating over the domain we end up with the exact solution u being a solution of the following weak form in the case of Dirichlet boundary conditions:
Find $u \in V$ such that

$$a(u, v) = (f, v) \quad v \in V, \tag{2.2.2}$$

where $a(\cdot,\cdot)$ represents the bilinear form which depends on L, $(\cdot,\cdot)$ represents an inner product and V being the Hilbert space. The Hilbert space has a weaker continuity requirements than is required in (2.2.1). We can also use this technique when we have a mix of boundary conditions with Dirichlet conditions on part of $\partial\Omega$ and Neumann boundary conditions on the rest of $\partial\Omega$ with a typical problem described in the form of finding u satisfying the following conditions. We let $\partial\Omega = \partial\Omega_D \cup \partial\Omega_N$ being the boundary of our domain Ω, with $\partial\Omega_D$ being the Dirichlet boundary part and $\partial\Omega_N$ being the Neumann

boundary part. Then our problem becomes

$$Lu = f \quad \text{in } \Omega, \tag{2.2.3}$$

$$u = \phi \quad \text{on } \partial\Omega_D, \quad \frac{\partial u}{\partial n} = g \quad \text{on } \partial\Omega_N, \tag{2.2.4}$$

with ϕ, g being suitable given functions. Now our extended weak form of the differential problem becomes as follows.
Find $u \in V$ such that

$$a(u, v) = (f, v) + < g, v > \quad v \in V, \tag{2.2.5}$$

where $< \cdot, \cdot >$ represents an inner product on the boundary with the precise details depending on the form of the operator L.

2.2.2 Abstract variational problems

We start by defining some terms with a general function space V and to be precise about the function spaces we take the following for one-dimensional problems with $(\cdot, \cdot)$ denoting an inner product and with $\|\cdot\|$ denoting the induced norm, i.e. $\|v\| = (v, v)^{1/2}$. An inequality that fairly quickly follows when we use this norm is the Cauchy-Schwartz inequality, see e.g. [27, p.77]:

$$|(u, v)| \leq \|u\| \, \|v\| \quad \forall u, v \in V. \tag{2.2.6}$$

Definition 1: ***The spaces*** $L_2(0,1)$ ***and*** $H^1(0,1)$
The spaces $L_2(0,1)$ and $H^1(0,1)$ are inner product spaces with inner products and norms as follows:

Space	*Inner product*	*Norm*
$L_2(0,1)$	$(u,v) = \int_0^1 uv dx$	$\|v\| = \left(\int_0^1 v^2 dx\right)^{\frac{1}{2}} < \infty$
$H^1(0,1)$	$(u,v) = \int_0^1 uv + u'v' dx$	$\|v\| = \left(\int_0^1 v^2 + v'^2 dx\right)^{\frac{1}{2}} < \infty$

Definition 2: *Bilinear Form*
We say $a : V \times V \to \mathbb{R}$ if

$$a(u, sv + tw) = sa(u, v) + ta(u, w), \quad a(su + tv, w) = sa(u, w) + ta(v, w)$$

for all $s, t \in \mathbb{R}$ and for all $u, v, w \in V$.

A bilinear form $a(\cdot, \cdot)$ in $V \times V$ is called symmetric if:

$$a(u, v) = a(v, u) \quad \forall u, v \in V \tag{2.2.7}$$

The bilinear form is **bounded** if there is a constant M such that

$$a(u, v) \leq M \|u\| \|v\| \quad \forall u, v \in V \tag{2.2.8}$$

The bilinear form is **coercive** if there exists a constant $c > 0$ such that

$$a(v, v) \geq c \|v\|^2 \quad \text{for all } v \in V \tag{2.2.9}$$

Remark
These spaces are examples of ***Hilbert spaces*** $(V, \|.\|)$. The finite element functions that we consider in one-dimensions when the region is $[0, 1]$ are continuous with the first derivative being piecewise continuous and are such that they are in $H^1(0, 1)$. In the context of expressing a problem in a weak form the following theorem is important.

Theorem: Lax-Milgram
Let V be a Hilbert Space. We consider $a : V \times V \to \mathbb{R}$ be a bounded bilinear form which is coercive. Then for every bounded linear form $F : V \to \mathbb{R}$ the weak problem:
Find $u \in V$

$$a(u, v) = F(v) \quad \forall v \in V \tag{2.2.10}$$

has a **unique** solution.
Proof and details can be found in [11].

When the bilinear form $a(\cdot, \cdot)$ is symmetric and positive definite on the Hilbert space V, i.e.

$$a(v, v) > 0, \quad \forall v \in V \quad \text{with } v \neq 0, \tag{2.2.11}$$

it defines an inner product on the space V known as the **energy inner product**. Therefore, we can define a norm associated with this inner product by:

$$|||v||| := a(v, v)^{\frac{1}{2}} \tag{2.2.12}$$

which is known as the **energy norm**.

2.3 Galerkin Method

The Galerkin method is a mathematical method used to obtain an approximate solution of the exact solution of partial differential equations. It uses a discretization process by which a mathematical problem is defined that can be solved on digital computers. In principle, it is the equivalent of applying the method of variation of parameters to a function space, by converting the equation to weak form. Typically, the f.e.m is based on the Galerkin method in which a finite set of basis functions are constructed using bases of continuous piecewise polynomial functions defined on **meshes**, of a finite number of non-overlapping **elements**. This thesis is based on the Galerkin method, which is used to find approximate solutions to the considered problems in two space dimensions, see chapter 3, and the simplified case of a one space dimensional problem, see chapter 6, and in chapter 6 we also consider a space-time problem.

2.3.1 Galerkin Orthogonality

The following remarks are based on [2], where you can find more analysis with the corresponding theorems and proofs for the f.e.m and the Galerkin approximation.

Let V_h denote a finite element subspace of the Hilbert space V where the weak problem is defined.

We consider the following weak problems: Find $u \in V$ and $u_h \in V$ such that

$$a(u, v) = F(v) \quad \forall v \in V \tag{2.3.1}$$

$$a(u_h, v) = F(v) \quad \forall v \in V_h \subset V \tag{2.3.2}$$

Then we subtract one equation from the other and we get:

$$a(u - u_h, v) = 0 \quad \forall v \ \in \ V_h \tag{2.3.3}$$

The equation (2.3.3) is known as **Galerkin orthogonality**. This means that the error $u - u_h$ in the Galerkin approximation is orthogonal to all functions in V_h in the energy inner product. Another way of saying this is that the approximation solution u_h is the **projection** of the exact solution u in the energy inner product.

Galerkin orthogonality is used in many places when we are considering the error $u - u_h$, e.g. in the lemma below, and it will appear later in the thesis when we consider the error

in a QoI and we consider the function spaces that can be used in our approximate dual problems. The best approximation property of the Galerkin approximation is given next.

Lemma 2.3.1: The approximation $u_h \in V_h$ is the best approximation to the exact solution u in the energy norm. Thus we have the following result:

$$|||u - u_h||| \leq |||u - v||| \quad \forall v \in V_h. \tag{2.3.4}$$

Proof

$$\begin{aligned} |||u - u_h|||^2 &= a(u - u_h, u - u_h) \\ &= a(u - u_h, u - u_h) + a(u - u_h, u_h - v) \\ &= a(u - u_h, u) - a(u - u_h, u_h) + a(u - u_h, u_h) - a(u - u_h, v) \\ &= a(u - u_h, u - v) \\ &\leq |||u - u_h|||||u - v||| \quad \forall v \in V_h \end{aligned}$$

where the last step is by Cauchy Schwartz inequality.

Remark
In particular, when $V_h \subset V$ is a space of piecewise polynomials defined with respect to a finite element mesh of Ω and $u_I \in V_h$ denotes an interpolant of u on the mesh then the best approximation property in the energy norm means that

$$|||u - u_h||| \leq |||u - u_I|||. \tag{2.3.5}$$

Results concerning how well polynomials interpolate functions on each element in a mesh can be used to show that $|||u - u_I||| \to 0$ as $h \to 0$ as the mesh is refined and hence $|||u - u_h||| \to 0$ as $h \to 0$.

2.4 Mathematical preliminaries for problems involving one variable

The previous subsection was mostly about a weak problem in general. In the case of approximating a function u by a piecewise polynomial function U we need suitable basis functions. We present next two possibilities for such basis functions when $u = u(x)$ depends on just one variable.

2.4.1 Lagrange polynomials as basis functions

Given a set of $n+1$ data points $(x_0, y_0), \ldots, (x_j, y_j), \ldots, (x_n, y_n)$ where no two x_j are the same, the Lagrange polynomial interpolation form is a *linear combination*

$$L(x) := \sum_{j=0}^{n} y_j l_j(x) \tag{2.4.1}$$

of Lagrange basis polynomials $l_0(x), \ldots, l_n(x)$ where

$$l_j(x) := \prod_{\substack{0 \le m \le n \\ m \ne j}} \frac{x - x_m}{x_j - x_m} = \frac{(x - x_0)}{(x_j - x_0)} \cdots \frac{(x - x_{j-1})}{(x_j - x_{j-1})} \cdot \frac{(x - x_{j+1})}{(x_j - x_{j+1})} \cdots \frac{(x - x_n)}{(x_j - x_n)}. \tag{2.4.2}$$

Each function $l_j(x)$ is a polynomial of degree n. By construction these functions have the property

$$l_j(x_i) = \begin{cases} 1, & i = j, \\ 0, & i \ne j \end{cases} \tag{2.4.3}$$

and hence $L(x_j) = y_j$.
Let $u \in C^{n+1}[a, b]$ and assume now that the distinct points $x_0, x_1, \ldots, x_n$ are in $[a, b]$. Let $L_n(x)$ be given by:

$$L_n(x) = \sum_{j=0}^{n} u(x_j) l_j(x). \tag{2.4.4}$$

From what is given above this is a polynomial of degree at most n which interpolates $u(x)$ at the $n + 1$ distinct points. It is the unique polynomial of degree at most n with this property as if two polynomials exist with this property then the difference between then is a polynomial of degree at most n with $n + 1$ distinct zeros and hence the difference is identically zero. The error in this approximation is defined by

$$e_n(x) = u(x) - L_n(x) \tag{2.4.5}$$

and the error can be expressed in the form

$$e_n(x) = \frac{1}{(k+1)!} u^{(n+1)}(c) \prod_{j=0}^{n} (x - x_j) \tag{2.4.6}$$

for some $c \in (a, b)$ with c depending on x and the $n + 1$ points.
Proof of the above error and more details can be found in [16, p.315].

In the context of a finite element computation and with an element being $[x_0, x_n]$ with

$x_0 < x_1 < \cdots < x_n$ the piecewise polynomial finite element basis functions can be defined in terms of the Lagrange polynomials $l_0(x), \ldots, l_n(x)$ when $x_0 \leq x \leq x_n$. The piecewise polynomial which corresponds to $l_j(x)$, $1 \leq j \leq n-1$ on $[x_0, x_n]$ is taken to be zero outside of the element and is thus continuous at the join points of x_0 and x_n. These piecewise polynomials are are hence only non-zero on one element. The piecewise polynomial which corresponds to $l_0(x)$ or $l_n(x)$ on $[x_0, x_n]$ is also non-zero in a neighbouring element.

2.4.2 Basis functions using Legendre Polynomials

Another way of representing the polynomial approximation on an element is to use a basis expressed in terms of Legendre polynomials and before we give this basis we first briefly introduce some important properties that we need about Legendre polynomials and, until we say otherwise, the interval involved is $[-1, 1]$.

There are a number of equivalent ways that Legendre polynomials can be defined with common ones being as follows. The Legendre polynomial $P_n(x)$ of degree n satisfies the Legendre differential equation

$$\frac{\mathrm{d}}{\mathrm{d}x}\left((1-x^2)\frac{\mathrm{d}}{\mathrm{d}x}P_n(x)\right) + n(n+1)P_n(x) = 0 \tag{2.4.7}$$

and it is given by Rodrigues formula

$$P_n(x) = \frac{1}{2^n n!}\frac{\mathrm{d}^n}{\mathrm{d}x^n}\left[(x^2-1)^n\right]. \tag{2.4.8}$$

The polynomials are also the coefficients in the Maclaurin series

$$\frac{1}{\sqrt{1-2xt+t^2}} = \sum_{n=0}^{\infty} P_n(x)t^n. \tag{2.4.9}$$

In the context of needing to be able to compute these polynomials the most convenient definition is the recursive definition

$$P_0(x) = 1, \quad P_1(x) = x \quad \text{and} \quad P_{k+1}(x) = \frac{(2k+1)xP_k(x) - kP_{k-1}(x)}{k+1}, \quad k = 1, 2, \ldots \tag{2.4.10}$$

An important property of the Legendre polynomials is that they are orthogonal with

respect to the L_2-norm on the interval $-1 \leq x \leq 1$, i.e.

$$\int_{-1}^{1} P_m(x)P_n(x)dx = \begin{cases} \frac{2}{2n+1}, & \text{when } m = n, \\ 0, & \text{otherwise.} \end{cases} \tag{2.4.11}$$

This property is used later in the thesis when higher order schemes in time are described to attempt to accurately solve the inflation of hyperelastic membranes under dynamic conditions. Another important property of Legendre polynomials is that they are odd or even, that is

$$P_n(-x) = (-1)^n P_n(x). \tag{2.4.12}$$

In addition, at the end points -1 and 1 we have

$$P_n(1) = 1 \quad P_n(-1) = (-1)^n. \tag{2.4.13}$$

This last property is useful in setting up basis functions which vanish at the ends of an interval as we can take as basis functions the following for an interval $-1 \leq s \leq 1$.

$$\begin{aligned} &b_0(s) = \frac{1-s}{2}, \quad b_n(s) = \frac{1+s}{2}, \\ &b_k(s) = P_{k+1}(s) - P_{k-1}(s), \quad k = 1, 2, \ldots, n-1. \end{aligned} \tag{2.4.14}$$

With this set-up the intermediate functions $b_1(s), \ldots, b_{n-1}(s)$ are 0 at $s = \pm 1$. In computations we typically also need to be able to evaluate the derivatives of the basis functions and there is another property of Legendre polynomials which can be used here which is that

$$b_k'(s) = P_{k+1}'(s) - P_{k-1}'(s) = (2k+1)P_k(s), \quad k \geq 1. \tag{2.4.15}$$

This property of Legendre polynomials follows quite quickly by using (2.4.9) and (2.4.10) as follows. By partially differentiating (2.4.9) with respect to x and also with respect to t gives respectively

$$t(1 - xt + t^2)^{-3/2} = \sum_{k=0}^{\infty} P_k'(x)t^k, \quad \text{i.e. } (1 - xt + t^2)^{-3/2} = \sum_{k=0}^{\infty} P_k'(x)t^{k-1},$$

and

$$(x - t)(1 - xt + t^2)^{-3/2} = \sum_{k=0}^{\infty} P_k(x)kt^{k-1}$$

and combining these last two results gives

$$(x - t)\sum_{k=0}^{\infty} P_k'(x)t^{k-1} = \sum_{k=0}^{\infty} P_k(x)kt^{k-1}.$$

Equating the coefficient of t^{k-1} gives

$$xP'_k(x) - P'_{k-1}(x) = kP_k(x). \tag{2.4.16}$$

Now if we differentiate the recurrence relation (2.4.10) then we also have

$$(k+1)P'_{k+1}(x) = (2k+1)(xP'_k(x) + P_k(x)) - kP'_{k-1}(x). \tag{2.4.17}$$

By using the expression for $xP'_k(x)$ from (2.4.16) in (2.4.17) and simplifying

$$\begin{aligned}(k+1)P'_{k+1}(x) &= (2k+1)(P'_{k-1}(x) + kP_k(x) + P_k(x)) - kP'_{k-1}(x) \\ &= (k+1)P'_{k-1}(x) + (2k+1)(k+1)P_k(x)\end{aligned}$$

By simplifying we get

$$P'_{k+1}(x) = P'_{k-1}(x) + (2k+1)P_k(x),$$

thus

$$P'_{k+1}(x) - P'_{k-1}(x) = (2k+1)P_k(x),$$

which gives the result (2.4.15).

Everything above has involved the standard interval $[-1, 1]$. In the case of an actual interval $[x_0, x_n]$ that we had before and a mapping

$$x : [-1, 1] \to [x_0, x_n], \quad x(s) = \left(\frac{x_0 + x_n}{2}\right) + \left(\frac{x_n - x_0}{2}\right) s \tag{2.4.18}$$

a polynomial of degree less than or equal to n can be expressed in the form

$$u(x(s)) = \sum_{k=0}^{n} c_k b_k(s)$$

with $c_0 = u(x_0)$ and $c_n = u(x_n)$ being the value of the function $u(x)$ at x_0 and x_n respectively. Assuming that the polynomial $u(x(s))$ is given is given in some way so that we can evaluate it at the end points to obtain c_0 and c_n as above there is more effort to

get the other coefficents when $n \geq 2$ as we next show. From the relations above

$$\frac{\mathrm{d}}{\mathrm{d}s}\left(u(x(s)) - c_0 b_0(s) - c_n b_n(s)\right) = \sum_{k=1}^{n-1} c_k b_k'(s) \tag{2.4.19}$$

$$= \sum_{k=1}^{n-1} c_k \left(P_{k+1}'(s) - P_{k-1}'(s)\right) \tag{2.4.20}$$

$$= \sum_{k=1}^{n-1} c_k (2k+1) P_k(s). \tag{2.4.21}$$

If we multiply this relation by $P_j(s)$ and integrate over $-1 \leq s \leq 1$ and use the orthogonality properties of the Legendre polynomials then we get

$$2c_j = \int_{-1}^{1} P_j(s) \frac{\mathrm{d}}{\mathrm{d}s}\left(u(x(s)) - c_0 b_0(s) - c_n b_n(s)\right) \mathrm{d}s. \tag{2.4.22}$$

2.5 Mathematical preliminaries for problems with two space dimensions

Now, we consider the case when our functions are of two space variables.

- **The divergence theorem**
  We define $\overline{\Omega} = \Omega \cup \partial\Omega$ where $\Omega \subset \mathbb{R}^2$, is a simply connected open bounded domain, with a piecewise smooth boundary $\partial\Omega$. Let $\underline{a} = (a_1, a_2)^T$ be a vector field with a_1 and a_2 being continuously differentiable in a domain containing $\Omega \cup \partial\Omega$ and let $\underline{n} = (n_1, n_2)^T$ be the unit outward normal vector to $\partial\Omega$. Then, the divergence theorem is defined as:

$$\iint_\Omega \nabla \cdot \underline{a} \quad dxdy = \int_{\partial\Omega} \underline{a} \cdot \underline{n} \quad ds \tag{2.5.1}$$

  where

$$\nabla \cdot \underline{a} = \frac{\partial a_1}{\partial x} + \frac{\partial a_2}{\partial y} \quad \text{and} \quad \underline{a} \cdot \underline{n} = a_1 n_1 + a_2 n_2$$

  and the line integral is in the positive sense.

- **The Green's Theorem**
  The Green's Theorem is a particular case of the divergence theorem when we take $\underline{a} = v\nabla u$ and states that

$$\iint_\Omega \nabla u \nabla v dxdy = \int_{\partial\Omega} v \frac{\partial u}{\partial n}\, ds - \iint_\Omega v \Delta u \, dxdy \tag{2.5.2}$$

where $u, v \in H^1(\overline{\Omega})$.

- **Transforming a double integral**
  We consider here how to transform an area integral, when we map from a standard element to a general element. In $\mathbb{R}^2$ the situation involves a one-to-one and onto mapping $\underline{x} : T \to \widetilde{T}$. We now consider the details concerning how we can change the region of integration from $\widetilde{T}$ in the (x, y) plane to the region T in the (s, t) plane. In order to change the region of integration from $\widetilde{T}$ in the (x, y) plane to the region T in the (s, t) plane we have to transform $(s, t) \to (x(s, t), y(s, t))$ by computing the absolute value of the determinant of the Jacobian matrix. The Jacobian matrix is given by

$$J = \begin{pmatrix} \frac{\partial x}{\partial s} & \frac{\partial x}{\partial t} \\ \frac{\partial y}{\partial s} & \frac{\partial y}{\partial t} \end{pmatrix} \tag{2.5.3}$$

  and its determinant by

$$\det(J) = \frac{\partial x}{\partial s}\frac{\partial y}{\partial t} - \frac{\partial x}{\partial t}\frac{\partial y}{\partial s}. \tag{2.5.4}$$

  Then, we get the transformation of a double integral by

$$\iint_{\widetilde{T}} f(x, y)dxdy = \iint_{T} f(x(s, t), y(s, t))|\det(J)|dsdt. \tag{2.5.5}$$

2.6 The Finite Elements Method in 2D

The following description of the f.e.m in 2D is used for the implementation of the general case of the membrane model in Chapter 3, which we refer to as the non-axisymmetric case.

2.6.1 Mesh in 2D

Let Ω denote a polygonal domain and let $\overline{\Omega} \equiv \Omega \cup \partial\Omega$ and we suppose that the region is partitioned into *ne* triangular elements which we refer to as a triangular mesh of Ω. In mathematical notation we have non-overlapping triangles $\overline{\Omega}_1, \overline{\Omega}_2, \ldots, \overline{\Omega}_{ne}$ so that:

$$\overline{\Omega} = \bigcup_{1}^{ne} \overline{\Omega}_i, \quad \Omega_i \cap \Omega_j = \emptyset, \ i \neq j \tag{2.6.1}$$

The individual triangles $\bar{\Omega}_1, \ldots, \bar{\Omega}_{ne}$ can have vertices and edges in common but they do not overlap.

The finite element spaces V_h are defined with respect to a mesh of the domain Ω of the problem.

Suppose that there are m vertices $\underline{\hat{x}}_1, \ldots, \underline{\hat{x}}_m$ in total in the mesh and to know which 3 vertices refer to which element we construct a matrix of size $3 \times ne$ with column r containing the 3 numbers i_{r1}, i_{r2}, i_{r3}which gives the 3 nodes associated with the r^{th} element. Therefore $\underline{\hat{x}}_{i_{r1}}, \underline{\hat{x}}_{i_{r2}}, \underline{\hat{x}}_{i_{r3}}$ are the 3 vertices describing the r^{th} triangle which we illustrate in figure 2.1. The matrix of size $3 \times ne$ gives what is known as the **connectivity** information. To illustrate what this involves the mesh shown in Figure 2.2 has the element/node connectivity information given in Table 2.6.1. The sequence of node numbers for any element can start with any node, and it helps at many parts if we arrange for the numbering to be such that the boundary is traversed in the counterclockwise direction although this can always be automatically arranged to be the case in a program.

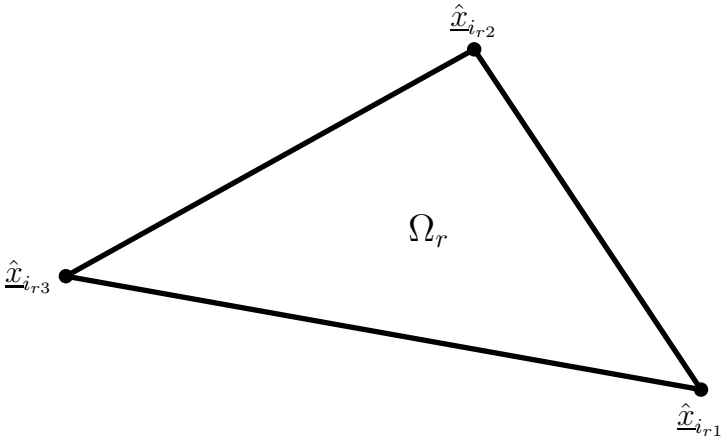


Fig. 2.1: The r^{th} triangle Ω_r in a mesh.

Tab. 2.6.1: Element/node table

ne		ne		ne	
1	(1,5,6)	7	(5,9,10)	13	(9,13,14)
2	(1,6,2)	8	(5,10,6)	14	(9,14,10)
3	(2,6,7)	9	(6,10,11)	15	(10,14,15)
4	(2,7,3)	10	(6,11,7)	16	(10,15,11)
5	(3,7,8)	11	(7,11,12)	17	(11,15,16)
6	(3,8,4)	12	(7,12,8)	18	(11,16,12)

In Figure 2.2 we can see an example of a uniform mesh of a rectangle with the node numbers and the element numbers displayed. We can see that each element number has 3 node numbers. It is also shows that each node is associated with a *patch* of elements. We can see, for example, that node 6 is a node of elements 2,1,8,9,10 and 3.

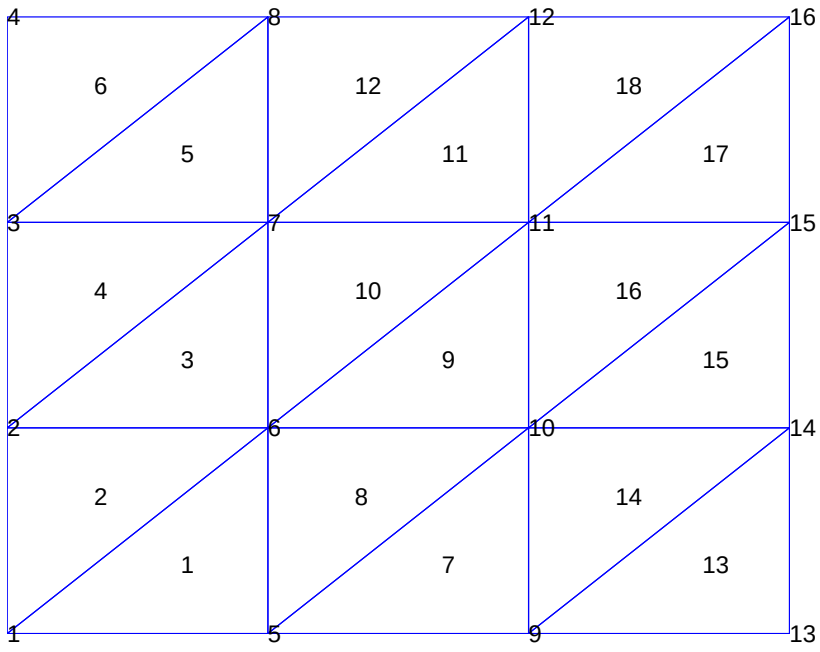


Fig. 2.2: A mesh with the nodes and elements

2.6.2 Implementation of the mesh

For the mesh implementation in Matlab, the data of the mesh is stored in two matrices with their name and shape indicated by

- **nodes(3,ne)**: which stores the 3 nodes of all the elements (ne).
  The 3 node numbers for the r^{th} element are given by *nodes(:, r)*.
- **coor(2,m)**: which stores the x and y coordinates of all the nodes of the mesh. The constant m denotes the total number of the nodes.
  The coordinates of the ith node are given by *coor(:,i)*.

When adaptive refinement is done we later describe briefly other matrices which can be determined from *nodes(.,.)* and *coor(.,.)* matrices which helps in the implementation of this process.

2.6.3 The finite element with piecewise linear functions

A surface plot of a piecewise linear basis function $\hat{\phi}_i(\underline{x})$ resembles a *pyramid* as we show in figure 2.3; that is they are piecewise linear functions that are linear in each element, satisfying:

$$\hat{\phi}_i(\underline{\hat{x}}_j) = \begin{cases} 1, & j = i, \\ 0, & j \neq i \end{cases} \quad \text{for } i, j = 1, \ldots, m \tag{2.6.2}$$

where $\underline{\hat{x}}_i$ are the nodes. With these functions the finite element space can be represented as:

$$V_h{}' = \text{span}\left\{\hat{\phi}_1, \ldots, \hat{\phi}_m\right\}. \tag{2.6.3}$$

Here, we have one basis function associated with each node in the mesh when we have one unknown parameter at each node, i.e. we can associate the function $\hat{\phi}_i$ with the unknown at the point $\underline{\hat{x}}_i$.

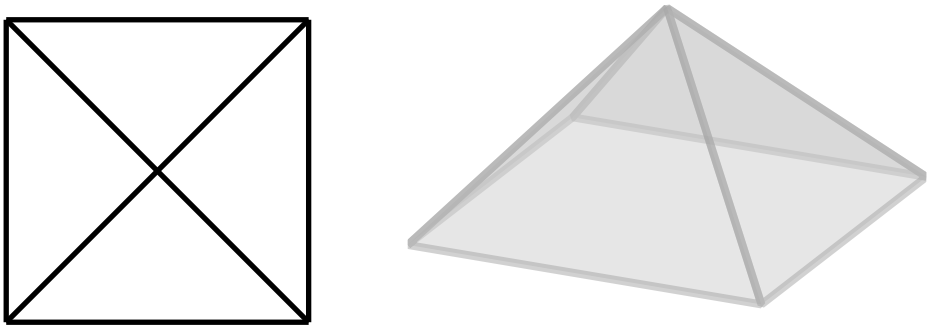


Fig. 2.3: Mesh of 4 triangles (left plot) and pyramid shape surface plot (right plot) of the piecewise linear basis function which is 1 at the centre node and 0 at the other 4 nodes.

Remark

For the definition of weak problem, which we use in our implementations, we use the subspace $\tilde{V}_h \subset V_h{}'$ which involves the test functions that vanish on the boundary $\partial\Omega$ on which there is a Dirichlet boundary condition. Therefore we define the following set

$$I_D = \{i : \quad \underline{x}_i \text{ is on the } \partial\Omega_D\}, \tag{2.6.4}$$

and note that $u_h(\underline{x}_i),\ i \in I_D$ is known. For the other values of i, i.e. for $i \in \{1, \ldots, m\}/I_D$ the values $u_h(\underline{x}_i)$ are not known at the start. Then the Galerkin method approximation

is of the form:

$$u_h(\underline{x}) = \sum_{i\in\{1,\ldots,m\}/I_D} u_h(\underline{x}_i)\hat{\phi}_i(\underline{x}) + \sum_{i\in I_D} u_h(\underline{x}_i)\hat{\phi}_i(\underline{x}) \tag{2.6.5}$$

and in the method we need to set-up equations to determine the unknown parameters.

2.6.4 The Galerkin method in 2D

As before, let Ω be a bounded domain in $\mathbb{R}^2$. We assume that in the two-dimensional case the boundary $\partial\Omega$ is a polygon, and that $\partial\Omega \equiv \partial\Omega_D \cup \partial\Omega_N$, with $\partial\Omega_D \cap \partial\Omega_N = \emptyset$ where $\partial\Omega_D$ and $\partial\Omega_N$ represent the Dirichlet and Neumann boundary condition respectively.
We define the subspace $\tilde{V}_h \subset V_h'$ as follows:

$$\tilde{V}_h = \{v(x,y) \in V_h' :\ v(x,y) = 0 \quad \forall (x,y) \in \partial\Omega_D\}. \tag{2.6.6}$$

Then, we consider the Poisson's equation

$$-\Delta u = -\nabla^2 u = -\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) = f(x,y) \quad \forall (x,y) \in \Omega \subset \mathbb{R}^2, \tag{2.6.7}$$

$$u(x,y) = \varphi(x,y)\ \forall (x,y) \in \partial\Omega_D, \quad \frac{\partial u}{\partial n}(x,y) = g(x,y)\ \forall (x,y) \in \partial\Omega_N. \tag{2.6.8}$$

where f, g_1, g_2 are given functions.
Now, we compute the variational form of this problem. By multiplying the equation (2.6.7) with a test function v, integrate over the region and then applying Green's Theorem (2.5.2) we have

$$\iint_\Omega f v dx dy = \iint_\Omega -v\Delta u dx dy = \iint_\Omega \nabla u \nabla v dx dy - \int_{\partial\Omega} v\frac{\partial u}{\partial n} ds \tag{2.6.9}$$

where $v \in \tilde{V}_h$.
Since $v(x,y) = 0\ \forall (x,y) \in \partial\Omega_D$, we have

$$\iint_\Omega \nabla u \nabla v dx dy = \iint_\Omega f v dx dy + \int_{\partial\Omega_N} v g ds. \tag{2.6.10}$$

The weak problem in two-dimensions, has the following form:
Find $\tilde{u}_h \in \tilde{V}_h$ such that

$$a(\tilde{u}_h, v) = (f, v) + < g, v > \quad \forall v \in \tilde{V}_h. \tag{2.6.11}$$

The Galerkin method approximation is of the form

$$\tilde{u}_h = \sum_{i=1}^{m} \tilde{u}_h(\underline{x}_i)\hat{\phi}_i. \tag{2.6.12}$$

The values at the nodes on the boundary $\partial\Omega_D$ are known and we need equations to determine the other values and these are given by

$$a(\tilde{u}_h, \hat{\phi}_i) = (f, \hat{\phi}_i) + < g, \hat{\phi}_i >, \quad i = \{1, \ldots, m\}/I_D. \tag{2.6.13}$$

If we substitute the expression for $\tilde{u}_h$ in (2.6.12) into (2.6.13) then for $i = \{1, \ldots, m\}/I_D$ we get that the unknown nodal parameters satisfy

$$\sum_{j \in \{1,\ldots,m\}/I_D} a(\hat{\phi}_j, \hat{\phi}_i)\tilde{u}_h(\underline{x}_j) = (f, \hat{\phi}_i) + < g, \hat{\phi}_i > - \sum_{j \in I_D} a(\hat{\phi}_j, \hat{\phi}_i)\tilde{u}_h(\underline{x}_j)$$

For the two-dimensional problems we need to do some extra computations, compared with one-dimensional problems. This happens because in this case we have to compute $a(\hat{\phi}_i, \hat{\phi}_j)$ and $(f, \hat{\phi}_i)$ for several elements instead of just 1 or 2 neighbouring elements as in the one-dimensional case. These computations are always organised in an element-by-element way as follows:
We define

$$a(u, v)_r = \iint_{\Omega_r} \nabla u \cdot \nabla v \, dxdy \tag{2.6.14}$$

$$(f, v)_r = \iint_{\Omega_r} f v \, dxdy \quad \text{and} \quad < g, v >_r = \int_{\Omega_{N_r}} gv \, ds \tag{2.6.15}$$

for all $r = 1, \ldots, ne$. Then we compute the sum of these elements contributions in order to compute $a(u, v)$,(f, v) and $< g, v >$. Thus we get

$$a(u, v) = \sum_{r=1}^{ne} a(u, v)_r, \quad (f, v) = \sum_{r=1}^{ne} (f, v)_r \quad \text{and} \quad < g, v > = \sum_{r=1}^{ne} < g, v >_r . \tag{2.6.16}$$

The element-by-element way of organizing things is to do all the calculations for each Ω_r and then put them all together according to Equation(2.6.16). The calculations to consider on element Ω_r involve all the basis functions which are non-zero on Ω_r. For this case, the corresponding situation involves a one-to-one and onto mapping $\underline{x} : T \to \Omega_r$, where T is the standard element in (s, t) plane and Ω_r the general element in (x, y) plane. Each point (s, t) is associated with a unique point $\underline{x}(s, t)$. The related basis functions

$\tilde{\phi}_i(\underline{x})$ defined on Ω_r have the following form:

$$\tilde{\phi}_i(\underline{x}(s,t)) = \phi_i(s,t) \quad i = 1, \ldots, me. \tag{2.6.17}$$

where me is the number of basis functions associated with element Ω_r and where $\phi_1, \ldots, \phi_{me}$ are the basis functions defined on T. In the case of linear triangles me=3 and we later consider 6-noded triangles when me=6.

2.6.5 Affine transformation

In this section we define the mapping $\underline{x} : T \to \Omega_r$, which is needed in order to compute the element matrix K_r and the element vector $\underline{b}_r$. Let T denote the right angled triangle which has vertices $(0,0), (1,0), (0,1)$ which we also refer to as $\underline{s}_1$, $\underline{s}_2$ and $\underline{s}_3$ respectively when that is convenient. Linear basis functions satisfying the Lagrange interpolation condition defined on this standard triangle are given by:

$$\phi_1(s,t) = 1 - s - t, \quad \phi_2(s,t) = s, \quad \phi_3(s,t) = t. \tag{2.6.18}$$

Let $u_{h_1} = u_h(\underline{x}_1), u_{h_2} = u_h(\underline{x}_2)$, and $u_{h_3} = u_h(\underline{x}_3)$. Then we have

$$\underline{x}(s,t) = \underline{x}_1\phi_1(s,t) + \underline{x}_2\phi_2(s,t) + \underline{x}_3\phi_3(s,t) \tag{2.6.19}$$

$$= \underline{x}_1 + (\underline{x}_2 - \underline{x}_1)s + (\underline{x}_3 - \underline{x}_1)t, \tag{2.6.20}$$

$$u_h(\underline{x}(s,t)) = u_{h_1}\phi_1(s,t) + u_{h_2}\phi_2(s,t) + u_{h_3}\phi_3(s,t) \tag{2.6.21}$$

$$= u_{h_1} + (u_{h_2} - u_{h_1})s + (u_{h_3} - u_{h_1})t. \tag{2.6.22}$$

We can observe that $\underline{s}_i \to \underline{x}_i, i = 1, 2, 3$, which means that the sides of T map to the sides of Ω_r and the interior of ∂T maps to the interior of $\partial\Omega_r$. This mapping form is known as an ***affine transformation***, that is a combination of a linear transformation followed by a translation. Now, we define the global functions $\tilde{\phi}_i,\ i = 1, 2, 3$ on the global triangle Ω_r by the relation

$$\tilde{\phi}_i(\underline{x}(s,t)) = \phi_i(s,t), \quad i = 1, 2, 3 \tag{2.6.23}$$

where ϕ_i denote the basis functions on the standard triangle T.

The element matrix

The element matrix (also commonly known as the element stiffness matrix) has the form:

$$K_r = \begin{pmatrix} a(\tilde{\phi}_1, \tilde{\phi}_1)_r & a(\tilde{\phi}_1, \tilde{\phi}_2)_r & a(\tilde{\phi}_1, \tilde{\phi}_3)_r \\ a(\tilde{\phi}_2, \tilde{\phi}_1)_r & a(\tilde{\phi}_2, \tilde{\phi}_2)_r & a(\tilde{\phi}_2, \tilde{\phi}_3)_r \\ a(\tilde{\phi}_3, \tilde{\phi}_1)_r & a(\tilde{\phi}_3, \tilde{\phi}_2)_r & a(\tilde{\phi}_3, \tilde{\phi}_3)_r \end{pmatrix} \tag{2.6.24}$$

In order to compute the element matrix, we need to find the integrals

$$a(\tilde{\phi}_i, \tilde{\phi}_j)_r = \iint_{\Omega_r} \left(\frac{\partial \tilde{\phi}_i}{\partial x} \frac{\partial \tilde{\phi}_j}{\partial x} + \frac{\partial \tilde{\phi}_i}{\partial y} \frac{\partial \tilde{\phi}_j}{\partial y} \right) dxdy, \quad i, j = 1, 2, 3. \tag{2.6.25}$$

We let

$$B = \begin{pmatrix} \frac{\partial \tilde{\phi}_1}{\partial x} & \frac{\partial \tilde{\phi}_2}{\partial x} & \frac{\partial \tilde{\phi}_3}{\partial x} \\ \frac{\partial \tilde{\phi}_1}{\partial y} & \frac{\partial \tilde{\phi}_2}{\partial y} & \frac{\partial \tilde{\phi}_3}{\partial y} \end{pmatrix} \tag{2.6.26}$$

which is constant on Ω_r, since we have linear basis functions (Equation(2.6.18)).
Then we can describe the element matrix by

$$K_r = (\text{area of } \Omega_r) B^T B. \tag{2.6.27}$$

Now we need to change the region of integration from the global element Ω_r in the (x, y) plane to the standard element T in the (s, t) plane. Therefore we have to transform $(s, t) \to (x(s, t), y(s, t))$ by computing the Jacobian matrix. Thus we have

$$J = \begin{pmatrix} \frac{\partial x}{\partial s} & \frac{\partial x}{\partial t} \\ \frac{\partial y}{\partial s} & \frac{\partial y}{\partial t} \end{pmatrix} = \begin{pmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{pmatrix} \tag{2.6.28}$$

where $(x_1, y_1), (x_2, y_2), (x_3, y_3)$ are the vertices of the global triangle Ω_r.
The magnitude $|\det(J)|$ of the determinant of J gives the ratio of an area increment in Ω_r to the corresponding increment in T and since J is constant on T and the area of T is $1/2$ we get that:

$$(\text{area of } \Omega_r) = \frac{|\det(J)|}{2} \tag{2.6.29}$$

Now, for the relation between the derivatives from the standard element to the actual element, by using the chain rule we get:

$$B = \begin{pmatrix} \frac{\partial \tilde{\phi}_1}{\partial x} & \frac{\partial \tilde{\phi}_2}{\partial x} & \frac{\partial \tilde{\phi}_3}{\partial x} \\ \frac{\partial \tilde{\phi}_1}{\partial y} & \frac{\partial \tilde{\phi}_2}{\partial y} & \frac{\partial \tilde{\phi}_3}{\partial y} \end{pmatrix} = J^{-T} \begin{pmatrix} \frac{\partial \phi_1}{\partial s} & \frac{\partial \phi_2}{\partial s} & \frac{\partial \phi_3}{\partial s} \\ \frac{\partial \phi_1}{\partial t} & \frac{\partial \phi_2}{\partial t} & \frac{\partial \phi_3}{\partial t} \end{pmatrix} = J^{-T} \begin{pmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix} \tag{2.6.30}$$

Therefore, the element matrix K_r has the form

$$K_r = \frac{|\det(J)|}{2} B^T B = \frac{|\det(J)|}{2} \begin{pmatrix} -1 & -1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} J^{-1} J^{-T} \begin{pmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}. \tag{2.6.31}$$

The element vector

The element vector (also commonly known as the load vector) has the form

$$\underline{b}_r = \begin{pmatrix} (f, \tilde{\phi}_1)_r \\ (f, \tilde{\phi}_2)_r \\ (f, \tilde{\phi}_3)_r \end{pmatrix}. \tag{2.6.32}$$

Therefore we have to compute the following integrals in the global region Ω_r:

$$(f, \tilde{\phi}_i)_r = \iint_{\Omega_r} f \tilde{\phi}_i \, dxdy, \quad i = 1, 2, 3. \tag{2.6.33}$$

In order to do these computations we first have to map from the region of integrations Ω_r to the standard triangle T and then use a quadrature rule to get an approximate solution. Thus we have

$$\underline{b}_r = \iint_T f(\underline{x}(s,t)) \begin{pmatrix} \phi_1(s,t) \\ \phi_2(s,t) \\ \phi_3(s,t) \end{pmatrix} |\det(J)| \, dsdt \tag{2.6.34}$$

$$= |\det(J)| \int_{t=0}^{t=1} \int_{s=0}^{s=1-t} f(\underline{x}(s,t)) \begin{pmatrix} 1-s-t \\ s \\ t \end{pmatrix} dsdt \tag{2.6.35}$$

$$\approx |\det(J)| \sum_{k=1}^{nq} w_k f(\underline{x}(s_k, t_k)) \begin{pmatrix} 1-s_k-t_k \\ s_k \\ t_k \end{pmatrix} \tag{2.6.36}$$

where $(s_k, t_k) \in T, k = 1, \ldots, nq$ are the quadrature points and $w_1, \ldots, w_{nq}$ are the weights. See Table 2.6.2 for the 1,3 and 7 point quadrature rules.

Tab. 2.6.2: Quadrature points and weights for integrating in $2D$

nq	Points	Weights
1	(1/3,1/3)	1/2
3	(1/2,0),(1/2,1/2),(0,1/2)	1/6,1/6,1/6
7	(0,0),(0.5,0),(1,0),(0.5,0.5), (0,1),(0,0.5),(1/3,1/3)	1/40, 1/15, 1/40, 1/15, 1/40, 1/15, 9/40

Neumann boundary condition

For the Neumann boundary condition, from the Equation (2.6.10), we have to compute the integrals

$$\int_{\partial\Omega_N} vg\, ds \tag{2.6.37}$$

for each of the basis functions v which are connected with an element which has an edge on $\partial\Omega_N$.

Suppose i_1, i_2 are the node numbers of an edge of a triangle which is on $\partial\Omega_N$. To ease the description suppose that all the nodes on the triangle are i_1, i_2, i_3 (in this order) in the anti-clockwise direction. In terms of the local basis functions, the edge corresponds to $t = 0$ for $0 < s < 1$, therefore we have

$$\phi_1(s,0) = 1 - s, \quad \phi_2(s,0) = s, \quad \phi_3(s,0) = 0. \tag{2.6.38}$$

The contribution from this edge is given by

$$\underline{b}_{i_1,i_2} = \int_0^1 g(\underline{x}(s,0)) \begin{pmatrix} \phi_1(s,0) \\ \phi_2(s,0) \end{pmatrix} \left| \frac{\partial \underline{x}(s,0)}{\partial s} \right| ds \tag{2.6.39}$$

$$= \left| \underline{x}_{i_2} - \underline{x}_{i_1} \right| \int_0^1 g(\underline{x}(s,0)) \begin{pmatrix} 1-s \\ s \end{pmatrix} ds. \tag{2.6.40}$$

To approximate these integrals using Gauss Legendre quadrature we transform the interval $0 \le s \le 1$ to the standard interval $-1 \le t \le 1$ for Gauss Legendre quadrature with $s(t) = (1+t)/2$ and this gives

$$\underline{b}_{i_1,i_2} = \left| \underline{x}_{i_2} - \underline{x}_{i_1} \right| \int_{-1}^1 g(\underline{x}(s(t),0)) \begin{pmatrix} \phi_1(s(t),0) \\ \phi_2(s(t),0) \end{pmatrix} dt \tag{2.6.41}$$

$$\approx \frac{\left| \underline{x}_{i_2} - \underline{x}_{i_1} \right|}{2} \sum_{k=1}^{nq} w_k g(\underline{x}(s_k,0)) \begin{pmatrix} 1-s_k \\ s_k \end{pmatrix} \tag{2.6.42}$$

where $s_k = s(t_k)$, and where t_k, $k = 1, \ldots, nq$ are the Gauss Legendre quadrature points and $w_1, \ldots, w_{nq}$ are the weights. For this case we used Gauss Legendre quadrature points and weights as given in Table 2.6.3. We note here that the above approximation in one space dimension by using Gauss Legendre quadrature in 1D is also used for the simplest axisymmetric membrane case which is described in Chapter 6.

Tab. 2.6.3: Quadrature points and weights for integrating in $1D$

nq	Points	Weights
2	$(1/\sqrt{3})$,(-$1/\sqrt{3}$)	1,1
3	(-$\sqrt{0.6}$),(0),($\sqrt{0.6}$)	5/9,8/9,5/9
4	$((-\sqrt{525+70\sqrt{30}})/35), ((-\sqrt{525-70\sqrt{3}})/35),$ $((\sqrt{525-70\sqrt{30}})/35), ((\sqrt{525+70\sqrt{30}})/35)$	$(18-\sqrt{30})/36, (18+\sqrt{30})/36,$ $(18+\sqrt{30})/36, (18-\sqrt{30})/36$
5	0 $\pm\frac{1}{21}\sqrt{245-14\sqrt{70}}$ $\pm\frac{1}{21}\sqrt{245+14\sqrt{70}}$	$\frac{128}{225}$ $\frac{1}{900}(322+13\sqrt{70}\ )$ $\frac{1}{900}(322-13\sqrt{70}\ )$

2.6.6 Assembling and solving the system

Global matrix $\hat{K}$

In order to compute the global stiffness matrix $\hat{K}$ of size $m \times m$ we have to sum up all the element matrices K_r for all $r = 1, \ldots, ne$ in an appropriate way. This process is called the assembly of $\hat{K}$.

Algorithm 1

$\hat{K}$ =zero matrix of size $m \times m$.

For $r = 1, \ldots,$ ne

$q =$ nodes$(:, r)$, the 3 node numbers of element r in matlab syntax

Compute the 3×3 element matrix K_r.

Replace $\hat{K}(q, q)$ by $\hat{K}(q, q) + K_r$.

End for loop

Remark

The global stiffness matrix $\hat{K}$ is a sparse matrix with the non-zero entries depending on how the nodes have been numbered and in an efficient implementation the matrix is stored in a sparse way. The entries on row i are all connected with the function $\hat{\phi}_i$ and this is non-zero only on the elements which have i as one of the nodes. It is the collection of all the nodes on all of these elements which gives the possible non-zero entries on row i of the $\hat{K}$ matrix.

Global vector $\underline{\hat{b}}$

Similar to the global matrix $\hat{K}$, in order to get the global load vector $\underline{\hat{b}}$ we have to sum up all the element vectors $\underline{b}_r$; for $r = 1, \ldots, ne$ in appropriate way. The assembling algorithm for the $\underline{\hat{b}}$ vector is the following

Algorithm 2

$\underline{\hat{b}}$ =zero column vector of length m.

For $r = 1, \cdots, \text{ne}$

$q = \text{nodes}(:, r)$, the 3 node numbers of element r

Find the nodes that are on the boundary,

by identifying by those that are on the Dirichlet or Neumann condition.

Compute the 3×1 element vector $\underline{b}_r$.

Compute the 3×1 element vector $\underline{b}_{N_r}$

for the Neumann boundary condition, if there is an edge on $\partial\Omega_N$.

Replace $\underline{\hat{b}}(q)$ by $\underline{\hat{b}}(q) + \underline{b}_r + \underline{b}_{N_r}$.

End for loop

Finally, in order to get the finite elements solution we have to solve the system $\hat{K}\underline{c} = \underline{\hat{b}}$

for the coefficients $\underline{c}$ of the finite element function u_h and in matlab syntax we have:

$$\underline{c} = \hat{K} \backslash \underline{\hat{b}} \tag{2.6.43}$$

2.6.7 The finite element method with piecewise quadratic functions

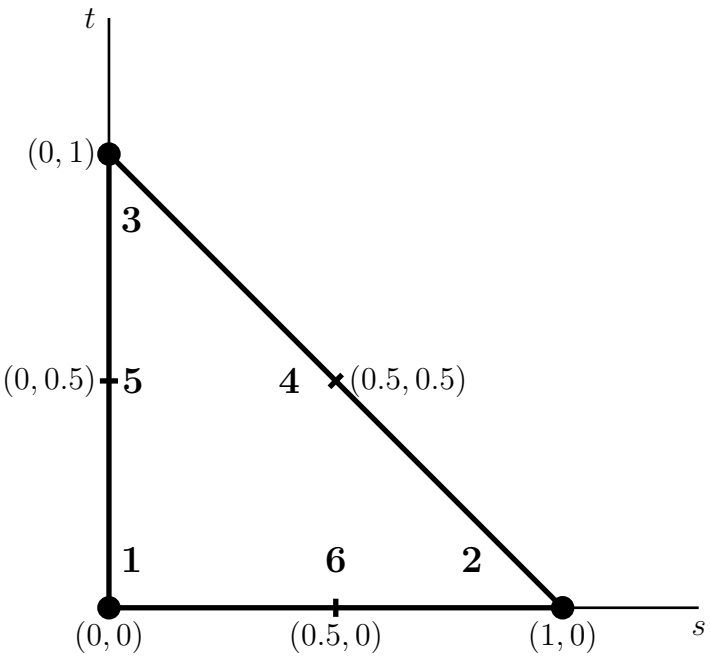
Later in this thesis we also use quadratic defined on triangles and in this case there are 6 nodes and 6 basis functions on each triangle. To describe these let Ω_r denote an actual element with vertices $\underline{x}_1$, $\underline{x}_2$ and $\underline{x}_3$ and let, as before,

$$\underline{x}(s,t) = \underline{x}_1 + (\underline{x}_2 - \underline{x}_1)s + (\underline{x}_3 - \underline{x}_1)t$$

be a mapping from the standard triangle to Ω_r. The 3 additional nodes are mid-side points and are given by

$$\underline{x}_4 = \frac{\underline{x}_2 + \underline{x}_3}{2}, \quad \underline{x}_5 = \frac{\underline{x}_3 + \underline{x}_1}{2} \quad \text{and} \quad \underline{x}_6 = \frac{\underline{x}_1 + \underline{x}_2}{2}.$$

With respect to the standard triangle the mid-side points are the images of $(0.5, 0.5)$, $(0, 0.5)$ and $(0.5, 0)$ with the standard triangles case shown below.



Using the standard triangle the 6 basis functions are given below.

$$\phi_1(\underline{x}(s,t)) = (1-s-t)(1-2s-2t). \tag{2.6.44}$$
$$\phi_2(\underline{x}(s,t)) = s(2s-1), \tag{2.6.45}$$
$$\phi_3(\underline{x}(s,t)) = t(2t-1), \tag{2.6.46}$$
$$\phi_4(\underline{x}(s,t)) = 4st, \tag{2.6.47}$$
$$\phi_5(\underline{x}(s,t)) = 4t(1-s-t), \tag{2.6.48}$$
$$\phi_6(\underline{x}(s,t)) = 4s(1-s-t). \tag{2.6.49}$$

The main difference in an implementation when quadratics are used is that the connectivity matrix is of size $6 \times ne$ with the element matrices being of size 6×6 and the element vector is of size 6×1.

2.7 A-Priori error Estimates for the Finite Element Method

In this section we present a few of the known error estimates for the finite element method and in all cases the term "error" usually means some norm of the error function e given

by

$$e := u - u_h \tag{2.7.1}$$

where, as defined earlier, u is the exact solution and u_h is the finite element approximation. There are two types of error estimates which are *a priori* and *a posteriori* estimates. The *a priori* estimates is based on analytical knowledge of the solution of the boundary value problem. The goal of these estimates is to give us a reasonable measure of the efficiency of a given method by telling us how fast the error decreases as we decrease the mesh size. In contrast, *a posteriori* estimates requires no *a priori* knowledge of the solution and is able to estimate the elemental errors in the mesh based on the finite element solution u_h. These estimates give us a much better idea of the actual error in a given finite element computation than do *a priori* estimates. Also, such estimates can be used to control **adaptive mesh refinement**. In adaptive mesh refinement, *a posteriori* error estimators are used to indicate where the error is particularly high, and more mesh intervals are then placed in those locations. A new finite element solution is computed, and the process is repeated until a satisfactory error tolerance is reached.

As this thesis is about estimating the error and improving the accuracy in approximations to functionals of u instead of to u itself we limit the detail to stating a few known results which influence how things are computed later and which help to justify the techniques that are done. *A'priori* estimates are discussed in this section and *a posteriori* error estimators are briefly mentioned in the next section where the emphasis at that stage is in describing the detail in constructing the adaptively refined mesh.

Many parts of the following are described in more detail in the book by Claes Johnson [15, Chapter 4].

By Lemma 2.3.1 u_h is the best approximation to u in the energy norm from the finite element space V_h of piecewise linear polynomials defined on a triangular mesh of Ω, i.e.

$$|||u - u_h||| < |||u - v||| \quad \forall v \in V_h. \tag{2.7.2}$$

Let $v = \pi_h u \in V_h$ be the nodal interpolant of u using the nodes in the triangular mesh. By estimating the interpolation error $|||u - \pi_h u|||$ we obtain a bound on the true error $|||u - u_h|||$.

Let πv denote the degree 1 polynomial interpolant to v on a triangle with the interpolation being at the nodal points.

To bound the energy norm of $v - \pi v$ on each element we need to consider both $v(x) - (\pi v)(x)$ and $\nabla v(x) - \nabla(\pi v)(x)$ on each triangle and the bounds depend on the size of the triangle and also on the angles of the triangle when the gradient is considered. For the size of the bounds to be small when the triangle is small we need to ensure that the mesh is such that no triangle Ω_r is arbitrarily thin which means that no angles can be allowed to be arbitrarily close to 0 or 180°. For the bounds we need the following quantities for Ω_r.

$$h_r = \text{the length of the longest side of } \Omega_r, \tag{2.7.3}$$

$$\rho_r = \text{the diameter of the largest circle which can be put in } \Omega_r. \tag{2.7.4}$$

Both h_r and ρ_r are illustrated in Figure 2.4.

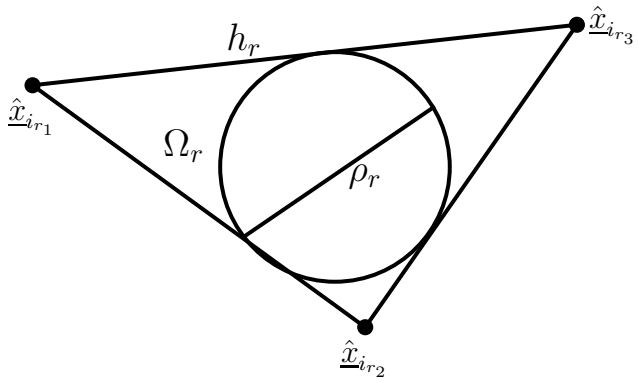


Fig. 2.4: Quantities for Ω_r

When the gradient is considered we need the meshes to be such that there exists a positive constant β, which is independent of h and which is such that

$$\frac{\rho_r}{h_r} \geq \beta \quad \text{for all triangles in the mesh.} \tag{2.7.5}$$

Meshes which satisfy this property are said to be quasi-uniform.

To obtain the bounds for a function $v \in C^2(\Omega_r)$ we need suitable representations for $v - \pi v$ and $\nabla v - \nabla(\pi v)$ and the details are a bit longer than in the much simpler one-dimensional case and are given in [15, Chapter 4]. Here we just state the results with comments added in some cases.

To avoid awkward notation involving double subscripts we let $\underline{a}_j = \underline{x}_{i_j}$, $j = 1, 2, 3$ when we consider the triangle Ω_r. A point (s, t) in the standard triangle gives the following

point $\underline{x}(s,t) \in \Omega_r$.

$$\underline{x}(s,t) = \underline{a}_1(1-s-t) + \underline{a}_2 s + \underline{a}_3 t = \underline{a}_1\tilde{\phi}_1(\underline{x}) + \underline{a}_2\tilde{\phi}_2(\underline{x}) + \underline{a}_3\tilde{\phi}_3(\underline{x}). \tag{2.7.6}$$

Also, recall that the 3 basis functions on Ω_r can be expressed in the form

$$\tilde{\phi}_1(\underline{x}(s,t)) = 1-s-t, \tag{2.7.7}$$
$$\tilde{\phi}_2(\underline{x}(s,t)) = s, \tag{2.7.8}$$
$$\tilde{\phi}_3(\underline{x}(s,t)) = t \tag{2.7.9}$$

and note in particular that for $\underline{x} \in \Omega_r$ we have $0 \leq \tilde{\phi}_j(\underline{x}) \leq 1$.

In the interpolation error results we need the following.

Lemma 2.7.1: Let $\tilde{\phi}_1$, $\tilde{\phi}_2$ and $\tilde{\phi}_3$ be as defined in (2.7.7)–(2.7.9) and let ρ_r be as defined in (2.7.4).

1.
$$\tilde{\phi}_1 + \tilde{\phi}_2 + \tilde{\phi}_3 = 1. \tag{2.7.10}$$

2.
$$\nabla(\tilde{\phi}_1 + \tilde{\phi}_2 + \tilde{\phi}_3) = \underline{0}. \tag{2.7.11}$$

3. For $i = 1, 2, 3$ and $j = 1, 2$ we have
$$\left|\frac{\partial\tilde{\phi}_i}{\partial x_j}\right| \leq \frac{1}{\rho_r}. \tag{2.7.12}$$

The first two results follow immediately from the definition of the 3 basis functions. To prove (2.7.12) you first need to note that all the first partial derivatives are constant on the triangle and if we take two points $\underline{p}$ and $\underline{q}$ in Ω_r then the ratio $|\tilde{\phi}_i(\underline{p}) - \tilde{\phi}_i(\underline{q})|/|\underline{p} - \underline{q}|$ is the same for any other two points on the same line. As we can take two points on the line to be at least a distance ρ_r part and as $|\tilde{\phi}_i(\underline{p}) - \tilde{\phi}_i(\underline{q})| \leq 1$ the bound follows.

The following two theorems, which use the previous lemma in their proofs, can alse be found in the same reference as the lemma and in particular see [15, p.85].

Theorem 2.7.1: Let Ω_r be a triangle with vertices $\underline{a}_1$, $\underline{a}_2$ and $\underline{a}_3$ in a quasi-uniform mesh with h_r and ρ_r defined in (2.7.3) and (2.7.4) respectively. Let $v \in C^2(\Omega_r)$ and let $\pi v \in$

span $\left\{\tilde{\phi}_1, \tilde{\phi}_2, \tilde{\phi}_3\right\}$ denote the interpolant to v with $\underline{a}_1$, $\underline{a}_2$ and $\underline{a}_3$ being the interpolation points. We have the following.

$$|v(\underline{x}) - \pi v(\underline{x})| \leq 2h_r^2 \max\left\{\left|\frac{\partial^2 v}{\partial x_i \partial x_j}\right| : \quad 1 \leq i, j, \leq 2\right\}, \tag{2.7.13}$$

$$\left|\frac{\partial v}{\partial x_k}(\underline{x}) - \frac{\partial \pi v}{\partial x_k}(\underline{x})\right| \leq 6\left(\frac{h_r^2}{\rho_r}\right) \max\left\{\left|\frac{\partial^2 v}{\partial x_i \partial x_j}\right| : \quad 1 \leq i, j, \leq 2\right\} \tag{2.7.14}$$

for $k = 1, 2$.

The theorem thus says that the pointwise error in the interpolant is $\mathcal{O}(h_r^2)$ and the pointwise error in the gradient is $\mathcal{O}(h_r)$ provided the triangle is such that (2.7.5) holds.

Theorem 2.7.2: Under the assumptions of Theorem 2.7.1 there is an absolute constant C such that

$$\|v - \pi v\|_{L_2(\Omega_r)} \leq Ch_r^2 |v|_{H^2(\Omega_r)}, \tag{2.7.15}$$

$$\|\nabla v - \nabla \pi v\|_{L_2(\Omega_r)} \leq C\frac{h_r^2}{\rho_r} |v|_{H^2(\Omega_r)}, \tag{2.7.16}$$

$$|v - \pi v|_{H^1(\Omega_r)} \leq C\frac{h_r^2}{\rho_r} |v|_{H^2(\Omega_r)}. \tag{2.7.17}$$

Note: $|.|_{H^r(\Omega)}$ denotes a *seminorm*, since we may have $|v|_{H^r(\Omega)} = 0$ even if $v \neq 0$.

These two Theorems 2.7.1 and 2.7.2 have exactly the same structure, the only difference being the norm involved, either L_∞ or the L_2-norm. Now, we are going to apply Theorem 2.7.2 to get the interpolation error estimates on the entire domain Ω in the case of $\|u - \pi_h u\|_{L_2(\Omega)}$, $\|\nabla u - \nabla \pi_h u\|_{L_2(\Omega)}$ and $|u - \pi_h u|_{H^1(\Omega)}$.
By summing the estimates above over all the triangles $\Omega_r \in T$ for each case we get the following:

- *for* $\|u - \pi_h u\|_{L_2(\Omega)}$:

$$\|u - \pi_h u\|^2_{L_2(\Omega)} = \sum_{\Omega_r \in T} \|u - \pi_h u\|^2_{L_2(\Omega_r)} \leq \sum_{\Omega_r \in T} C^2 h_r^4 |u|^2_{H^2(\Omega_r)} \tag{2.7.18}$$

$$\leq C^2 h^4 \sum_{\Omega_r \in T} |u|^2_{H^2(\Omega_r)} = C^2 h^4 |u|^2_{H^2(\Omega)}. \tag{2.7.19}$$

Therefore we get

$$\|u - \pi_h u\|_{L_2(\Omega)} \leq Ch^2 |u|_{H^2(\Omega)} \tag{2.7.20}$$

- *for* $\|\nabla u - \nabla \pi_h u\|_{L_2(\Omega)}$:
  Similarly by using $\frac{h_r}{\rho_r} \leq \frac{1}{\beta}$, by (2.7.5) we get

$$\|\nabla u - \nabla \pi_h u\|^2_{L_2(\Omega)} \leq \sum_{\Omega_r \in T} C^2 \frac{h_r^4}{\rho_r^2} |u|^2_{H^2(\Omega_r)} \leq \sum_{\Omega_r \in T} \frac{C^2 h_r^2}{\beta^2} |u|^2_{H^2(\Omega_r)} \tag{2.7.21}$$

$$\leq \frac{C^2 h^2}{\beta^2} |u|^2_{H^2(\Omega)} \tag{2.7.22}$$

  so that

$$\|\nabla u - \nabla \pi_h u\|_{L_2(\Omega)} \leq \frac{Ch}{\beta} |u|_{H^2(\Omega)}. \tag{2.7.23}$$

  The bound is of the form

$$Ch|u|_{H^2(\Omega)}$$

  if the constant C is redefined to include β.

- *for* $|u - \pi_h u|_{H^1(\Omega)}$:
  Similarly to the previous case we have:

$$|u - \pi_h u|^2_{H^1(\Omega)} \leq \sum_{\Omega_r \in T} C^2 \frac{h_r^4}{\rho_r^2} |u|^2_{H^2(\Omega_r)} \leq \sum_{\Omega_r \in T} \frac{C^2 h_r^2}{\beta^2} |u|^2_{H^2(\Omega_r} \tag{2.7.24}$$

$$\leq \frac{C^2 h^2}{\beta^2} |u|^2_{H^2(\Omega)} \tag{2.7.25}$$

  therefore

$$|u - \pi_h u|_{H^1(\Omega)} \leq \frac{Ch}{\beta} |u|_{H^2(\Omega)}. \tag{2.7.26}$$

  The bound is of the form

$$Ch|u|_{H^2(\Omega)},$$

  if the constant C is redefined to include β.

Remarks

(i) We can observe from the results above, that all bounds of global interpolation errors depend on the second partial derivatives of the exact solution u, on the constant C and on the mesh size h.

(ii) In principle a bound on C can be determined but $|u|_{H^2(\Omega_k)}$ is not known and here we cannot compute the quantities in (2.7.20), (2.7.23) or (2.7.26) to drive a refinement procedure.

2.8 A-posteriori error indicators and adaptive mesh refinement

There are many *a-posteriori* error indicators described in the book by Ainsworth and Oden [1] and in the book by Babuška, Whiteman and Strouboulis [5]. For the purpose of this section we just describe how to compute one of these which is due to Bank and Weiser [7] which gives us quantities to drive an adaptive refinement procedure which we describe. We need a problem to consider which we take here as the following Poisson problem.

Let Ω be a polygonal domain with boundary $\partial\Omega$ and assume that the set-up is such that there is a unique solution u of the Poisson problem given in (2.6.7)-(2.6.8) which we repeat here in slightly abbreviated form as

$$-\Delta u = f \quad \text{in } \Omega, \tag{2.8.1}$$

$$u = \phi \quad \text{on } \partial\Omega_D, \qquad \frac{\partial u}{\partial n} = g \quad \text{on } \partial\Omega_N, \tag{2.8.2}$$

where $\partial\Omega = \partial\Omega_D \cup \partial\Omega_N$ is a partition of $\partial\Omega$ and where f, ϕ and g are suitable given functions.

2.8.1 Computing one of the Bank Weiser error indicators

There are several error indicators described in [7] and we just consider one of these here. Roughly, the indicators are derived in order to attempt to solve for the error or more precisely with quantities which are consistent with the error. To describe the estimator that we compute we need the following function spaces for functions defined on a triangle Ω_r.

$$\bar{V}_{\Omega_r} = \{v : \quad v \text{ is a degree} \leq 2 \text{ polynomial on } \Omega_r\}, \tag{2.8.3}$$

$$\breve{V}_{\Omega_r} = \left\{v \in \bar{V}_{\Omega_r} : \quad v = 0 \text{ at the 3 nodes of } \Omega_r\right\}. \tag{2.8.4}$$

On the triangle Ω_r we have a basis for $\breve{V}_{\Omega_r}$ by using 3 of the basis functions given in section 2.6.7, i.e.

$$\breve{V}_{\Omega_r} = \text{span}\left\{\phi_4, \phi_5, \phi_6\right\}.$$

Let u_h denote the piecewise linear finite element solution of (2.8.1)-(2.8.2), as before, let $e = u - u_h$ denote the error. One of the quantities that we need to compute for functions

$v \in \breve{V}_{\Omega_r}$ is the following

$$F_r(v) := \iint_{\Omega_r} fv\, dx_1 dx_2 + \int_{\partial\Omega_r \cap \partial\Omega_N} v\left(g - \frac{\partial u_h}{\partial n}\right) \mathrm{d}s + \frac{1}{2}\int_{\partial\Omega_r/\partial\Omega} v\left[\frac{\partial u_h}{\partial n}\right]_J \mathrm{d}s \quad (2.8.5)$$

where

$$\left[\frac{\partial u_h}{\partial n}\right]_J := \left(\nabla u_h|_{\Omega_{r'}} - \nabla u_h|_{\Omega_r}\right) \cdot \underline{n}_r \quad (2.8.6)$$

is the jump in the normal derivative of u_h across an edge in the mesh. (In this description $\Omega_{r'}$ denotes the triangle on the other side of an edge to Ω_r.) Also for functions $\breve{e}$ and v in $\breve{V}_{\Omega_r}$ let

$$a(\breve{e}, v)_r = \iint_{\Omega_r} \nabla\breve{e} \cdot \nabla v\, dx_1 dx_2. \quad (2.8.7)$$

The Bank Weiser estimator that we compute is defined as the function $\breve{e} \in \breve{V}_{\Omega_r}$ such that

$$a(\breve{e}, v)_r = F_r(v) \quad \forall v \in \breve{V}_{\Omega_r}. \quad (2.8.8)$$

With the basis that we have for $\breve{V}_{\Omega_r}$ we have

$$\breve{e} = c_4\phi_4 + c_5\phi_5 + c_6\phi_6 \quad (2.8.9)$$

where c_4, c_5 and c_6 satisfy the equations

$$\breve{K}\begin{pmatrix} c_4 \\ c_5 \\ c_6 \end{pmatrix} = \begin{pmatrix} F_r(\phi_4) \\ F_r(\phi_5) \\ F_r(\phi_6) \end{pmatrix}, \quad \text{where } \breve{K} = \begin{pmatrix} a(\phi_4,\phi_4)_r & a(\phi_4,\phi_5)_r & a(\phi_4,\phi_6)_r \\ a(\phi_5,\phi_4)_r & a(\phi_5,\phi_5)_r & a(\phi_5,\phi_6)_r \\ a(\phi_6,\phi_4)_r & a(\phi_6,\phi_5)_r & a(\phi_6,\phi_6)_r \end{pmatrix}. \quad (2.8.10)$$

Once we have $\underline{c} = (c_4,\, c_5,\, c_6)^T$ we compute

$$a(\breve{e}, \breve{e})_r = \underline{c}^T \breve{K} \underline{c} \quad (2.8.11)$$

as the estimate of $a(e, e)_r$.

2.8.2 Adaptive Mesh Refinement

General discussion

The previous section described how to compute an error indicator and in this section we describe a strategy in which we use it as the basis for our refinement decision to attempt

to get a desired accuracy. We only consider h-refinement here. When a triangle is marked for refinement it is replaced by 4 similar triangles in the next mesh. The knock-on effect of this is that sides of the neighbouring triangles also need refining. If this gives triangles with just 2 sides which need refining then such a triangle is marked for full refinement which can have a further knock-on effect of more neighbouring triangles also needing some refinement. As a consequence a small number of steps is typically needed before a situation is reached where the refinement decision on every triangle is one of the following.

- The triangle is to be divided into 4 similar triangles which involves 3 new nodes at the mid-point on each side.
- The triangle is to be divided into 2 triangles which involves 1 new node at the mid-point of one of the sides. A triangle in this category is next to exactly one triangle which needs refining and it is sometimes referred to as a transition triangle. A subdivision of a triangle into two parts is done provided no new angle is created which is too small and if the division of the triangle in this way would cause this then it is instead divided into 4 similar triangles.

The set-up is such that there are no hanging nodes in the new mesh. We illustrate the two cases in figures 2.5 and 2.6 and we shortly give some information about the bookkeeping needed to do refinement in this way.

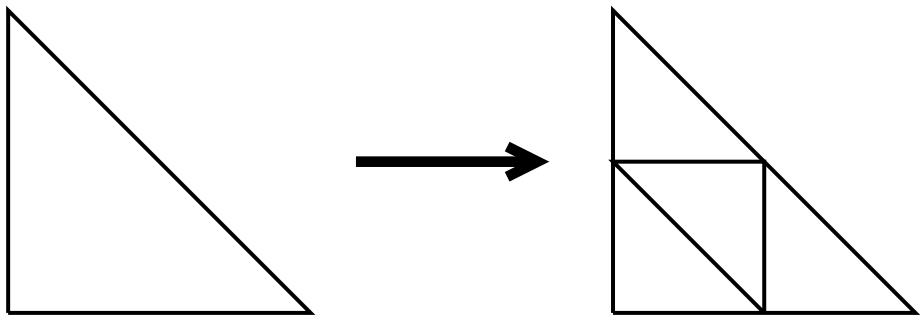


Fig. 2.5: A triangle which is fully refined into 4 similar triangles

Which elements to refine?

Before the bookkeeping part is given we describe an algorithm which uses the estimators $a(\breve{e}, \breve{e})_r$, $r = 1, \cdots, ne$ to attempt to get a given accuracy in the energy norm. If $\epsilon > 0$ and the aim is to attempt to get an approximation u_h such that in the energy norm

$$|||u - u_h||| < \epsilon \tag{2.8.12}$$

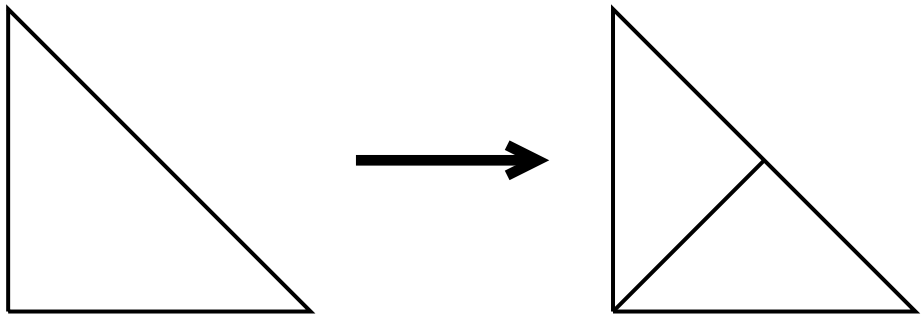


Fig. 2.6: A triangle which is divided into 2 triangles

then we compute until our estimate of the error in the energy norm satisfies

$$\sum_{r=1}^{ne} a(\breve{e}, \breve{e})_r < \epsilon^2. \tag{2.8.13}$$

A possible algorithm is as follows.

Algorithm 3

1. With an initial mesh, which is not too coarse, we calculate the finite element solution which we denote by u_h^{old}.
2. We compute the error estimate $|||\breve{e}|||_{\Omega_r}$ on each element.
3. We compute our estimate of $|||u - u_h|||$ and stop if the estimate satisfies (2.8.13).
4. We mark for refinement all triangles for which $a(\breve{e}, \breve{e})_r$ is 'large' and we create a new mesh. We discuss what large means in a moment.
5. Using the new mesh we calculate the finite element approximation which we denote by u_h^{new} and we replace u_h^{old} with u_h^{new}.
6. Repeat items 2-6 until the condition in item 3 has been satisfied or we have reached a bound on the number of elements that we will consider.

In item 4 there are different possibilities for deciding that the contribution of the r^{th} element to the global error is too large. Now if the mesh and the approximation u_h are such that

$$|||\breve{e}|||^2_{\Omega_r} \leq \frac{\epsilon^2}{ne}, \quad \text{for } r = 1, \ldots, ne, \tag{2.8.14}$$

then, by summing up over all triangles we get

$$|||\breve{e}|||_{\Omega}^2 = \sum_{r=1}^{ne} |||\breve{e}|||_{\Omega_r}^2 \leq \epsilon^2 \tag{2.8.15}$$

and the decision in item 3 is to stop the calculation. Hence one criteria for marking a triangle for refinement is to take all triangles Ω_r for which

$$|||\breve{e}|||_{\Omega_r}^2 > \frac{\epsilon^2}{ne}. \tag{2.8.16}$$

Comments on the bookkeeping in the refinement

With `ne` triangles involving `m` nodal points we have already mentioned in section 2.6.2 that the data about the mesh is stored in a Matlab program in matrices `nodes` and `coor` which are of size $3 \times$ `ne` and $2 \times$ `m` respectively and we write them as `nodes(3,ne)` and `coor(2,m)`. From `ne`, `m`, `nodes` and `coor` we can generate other quantities and these are all listed in the following table.

`ne`	Number of elements.
`m`	Number of nodes.
`ns`	Number of sides.
`nodes(3,ne)`	The `r`th column contains the 3 node numbers for the `r`th triangle. If we denote here the numbers as `i1`, `i2` and `i3` then it is convenient to arrange these so that the closed path `i1` to `i2` to `i3` to `i1` is anti-clockwise.
`coor(2,m)`	The `i`th column contains the x and y coordinates of the `i`th node.
`sides_info(4, ns)`	In the `i`th column there are 4 numbers with the first two numbers being the 2 node numbers of the side and the last two numbers being the element numbers for which this is a side. If the side is on the edge of the domain then the 4th entry is set to -1 to indicate this.
`sid_el(3, ne)`	The `r`th column of this gives the 3 side numbers associated with the `r`th triangle. If these 3 numbers are denoted by `j1`, `j2` and `j3` and as shown in figure 2.7 and if `i1`, `i2` and `i3` denote the 3 node numbers for this element then the numbers are arranged so that side `j1` is `i1` to `i2`, side `j2` is `i2` to `i3`, and side `j3` is `i3` to `i1`.
`ibc(m)`	This is not needed to describe the mesh but is useful to mention here. The `i`th entry is set to 0 for an interior node, it is set to 1 for a node on the boundary where there is a Dirichlet boundary condition and it is set to 2 on the boundary where there is a Neumann boundary condition.

Briefly, to obtain `ns` and `sides_info(4, ns)` from nodes we loop through the `ne` elements to construct an intermediate matrix of size $3 \times (3\texttt{ne})$ with the `r`th triangle contributing 3 columns of the following form.

```
min([i1, i2]),  min([i2, i3]),  min([i3, i1]),
max([i1, i2]),  max([i2, i3]),  max([i3, i1]),
      r,              r,              r
```

Then with sorting the entries from all the triangles we determine the number of different sides `ns` and we get the 1 or 2 triangles associated with each side to create `sides_info(.,.)`. The matrix `sid_el(.,.)` is then generated by looping through the columns of `sides_info(.,.)` to collect the side numbers for each triangle. The remaining

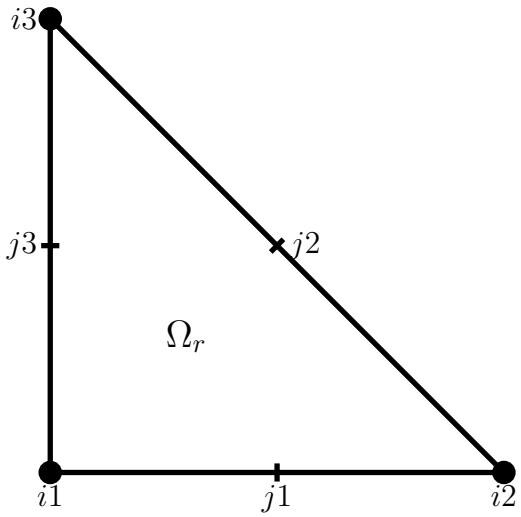


Fig. 2.7: Nodes and sides of the `r`th triangle with the 3 nodes being stored as `nodes(:, r)` and with the 3 side numbers being stored as `sid_el(:, r)`.

operations are concerned with tidying-up so that the quantities in each column are in the order indicated in the table.

With the extra information above we have all the information to quickly determine for each triangle which triangle shares each edge which is needed when we consider the jump in the gradient vector of the approximate solution from one triangle to its neighbouring triangles when we compute the error estimator. We also have the information needed to be able to set-up the matrices `nodes` and `coor` in the next finer mesh. When a triangle is marked for refinement all 3 sides are marked for refinement and we iterate, if necessary, until we get to a state that each triangle needs either 1 side or all 3 sides to be refined. The vertices in the finer mesh which were not in the coarser mesh and are the mid-points of all the sides which are being refined.

2.9 Numerical Results

For the implementation of the finite element method in 2D, we used the following problem. We let the square domain

$$\Omega = \{(x, y) : \quad |x| < 1,\ |y| < 1\}\,. \tag{2.9.1}$$

For the boundary $\partial\Omega = \partial\Omega_D \cup \partial\Omega_N$ we use the following: For the Neumann boundary condition $\partial\Omega_N$ we used:

$$\partial\Omega_N = \{(-1, y) : |y| < 1\} \tag{2.9.2}$$

therefore, for the Dirichlet boundary condition we have:

$$\partial\Omega_D = \partial\Omega \backslash \partial\Omega_N. \tag{2.9.3}$$

$$-\Delta u = f \quad \text{in } \Omega \tag{2.9.4}$$

$$u = x^4 + y^4 \quad \text{on } \partial\Omega_D \tag{2.9.5}$$

$$\frac{\partial u}{\partial n} = g \quad \text{on } \partial\Omega_N \tag{2.9.6}$$

where $f = 12(x^2 + y^2)$ and g is set so that $u = x^4 + y^4$ is the solution everywhere for all choices of the domain Ω.

In the computations we use the Bank Weiser estimator $\check{e}$ for the mesh refinement decisions.

In Figures 2.8–2.11 we show 4 meshes with the adaptive refinement done via the calculation of $\check{e}$ with computation done until the mesh and the approximation are such that $|||\check{e}||| < 10^{-2}$. In table 2.9.4 we show the estimates that are obtained and we also compare with the exact error which we can compute in this example as we know the exact solution. The values in the last column of the table correspond to

$$\frac{|||\check{e}|||}{|||u - u_h|||}. \tag{2.9.7}$$

The table only suggests that the estimator is consistent with the true error which is all that is shown in the theory. The refinement around the edge of Ω which are the parts furthest from $(0, 0)$ would have been predicted as this is where the second partial derivatives of u are largest in magnitude. By looking the figures we can observe that on the left hand side edge, see Figure 2.11, we have more refined elements. The additional refinement on this side edge is because this is where we have a Neumann boundary condition where we are estimating u whereas on the other sides u is known as it is given by the Dirichlet boundary condition.

Tab. 2.9.4: Numerical Results

ne	m	$\|\|\|u - u_h\|\|\|$	$\|\|\|\breve{e}\|\|\|$	ratio
162	100	9.757581e-002	2.852447e-001	2.923314e+000
550	312	1.278624e-002	5.546608e-002	4.337952e+000
1042	594	6.250652e-003	1.901255e-002	3.041690e+000
1343	763	5.815998e-003	9.232490e-003	1.587430e+000

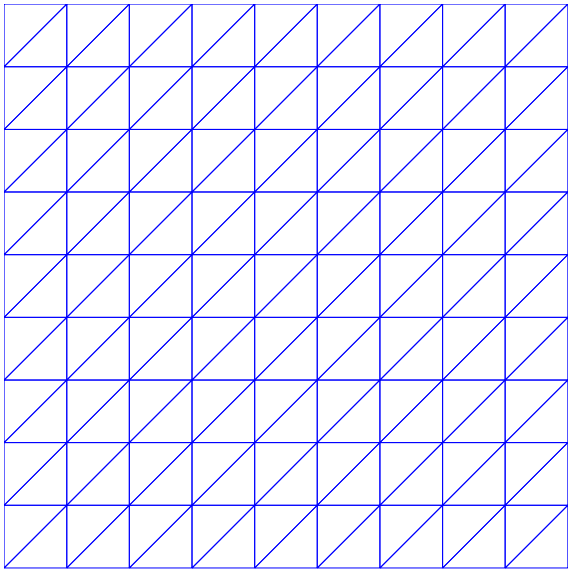


Fig. 2.8: Starting Mesh with `ne`=162, `m`=100.

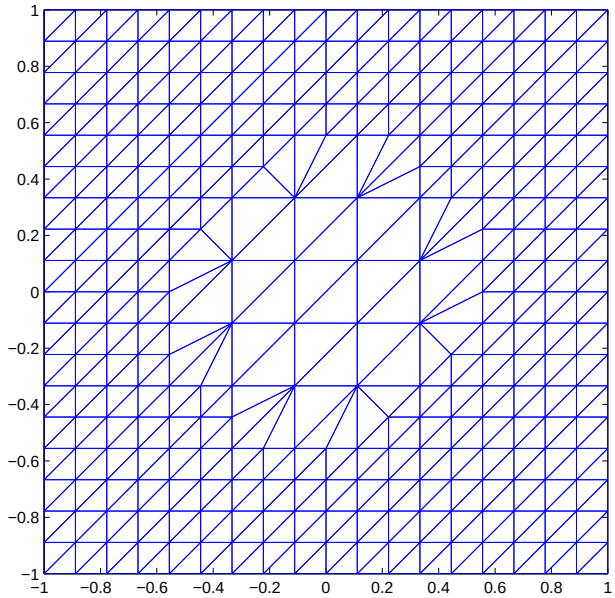


Fig. 2.9: The next refinement mesh with `ne`=550, `m`=312.

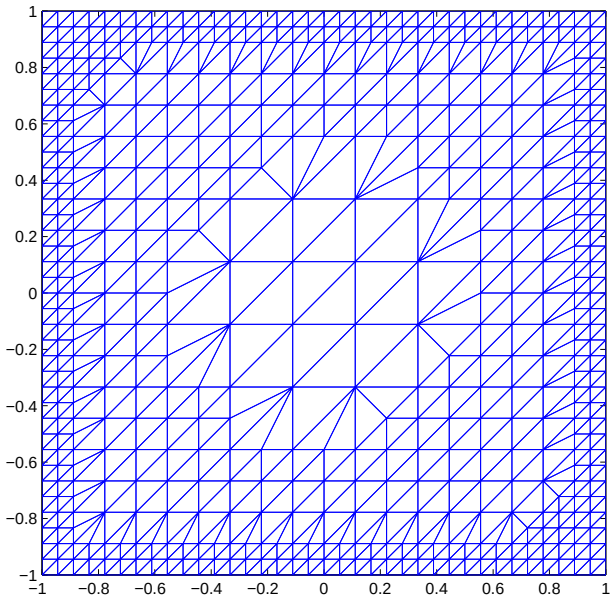


Fig. 2.10: The next refinement mesh with `ne`=1042, `m`=594.

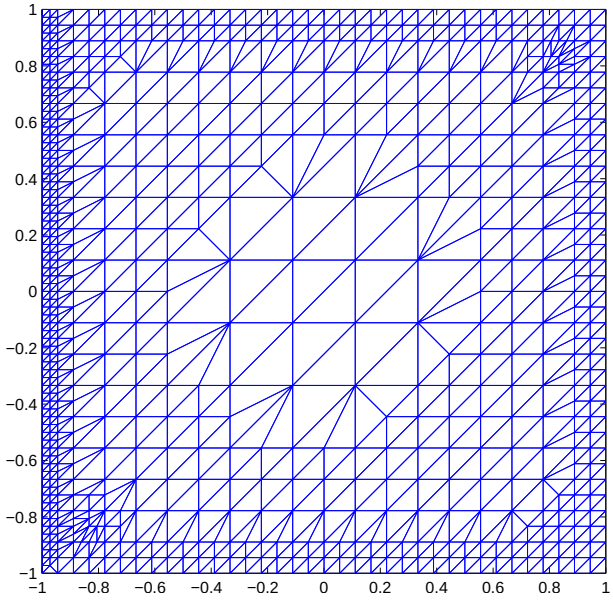


Fig. 2.11: The last mesh considered with `ne`=1343, `m`=763

3. THE MEMBRANE MODEL FOR A QUASI-STATIC DEFORMATION

3.1 Introduction

In this chapter we will give a full description of the finite inflation of a thin sheet modelled as a hyperelastic membrane. This is a non-linear problem, where the computational modelling of this involves finding, by using f.e.m., approximations to the displacement field, denoted by $\underline{u}$, of the deformed membrane under a given pressure loading. Then, these f.e. approximations are used in order to estimate some quantities of physical interest derived from the solution $\underline{u}$ by applying a functional J. For this case we consider as quantities of interest (i) the averaged thickness stretch ratio of the membrane over part of the domain and (ii) the potential energy of the deformed membrane.

As was described in the previous chapter, in order to get an estimate of the error in the given functional J, we set-up and solve a so-called **dual problem**, which is related to the given QoI and the weak problem $A(\cdot;\cdot)$ which describes the membrane model. This is done near the end of the chapter.

First we start with the description of a general 3D solid in order to get the weak form of the equations of equilibrium. Later, in section 3.4, we describe the membrane model for how the thin sheet deforms and in particular give the weak form of the equations of equilibrium under pressure loading in this simplified case. To complete the description of the mathematical model we give in section 3.5 some standard hyperelastic constitutive relations for the incompressible case and in all cases these hyperelastic models are expressed in the Ogden form [21], which we use in all our implementations in this thesis.

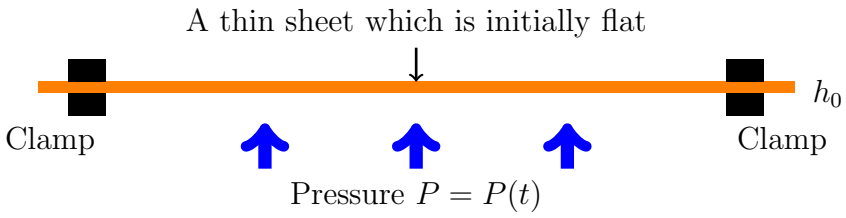
The final sections of the chapter are concerned with describing the numerical scheme and some aspects of the finite element implementation in order to solve the nonlinear problem for the membrane case under pressure. For the implementation of the problem

we used Newton's method for each system of nonlinear equations encountered in order to get an approximate solution $\underline{u_h}$ for the displacement $\underline{u}$. With an approximation $\underline{u_h}$ to $\underline{u}$ obtained we then move to the goal-oriented error estimation, where the error is measured with respect to a specific quantity of interest. Our numerical scheme for the error in the given QoI extends a little what is given in [28] to the non-axisymmetric situation described in this chapter. In section 4.6 we describe the **dual problem** which is based on the Gâteaux derivatives of the expression A in the weak problem (see later) and the given functional J which describes our QoI. In this chapter we restrict the detail to the expressions involved with examples of using the technique and various numerical results given in the next chapter.

3.2 The membrane assumptions and the weak form

In this section we consider the assumptions used in a membrane model of a thin sheet and we introduce the expressions in the weak form of the problem with the details of their derivation done in later sections.

The situation at the start of the process involves a thin sheet of uniform thickness h_0 which is initially flat and which is clamped at its edge as indicated below. This is the undeformed state of the body. The thin sheet deforms when a pressure is applied.



For later reference we let $\underline{N} =$ be the direction normal to the sheet at this stage. 2e assume that the body is composed of homogeneous, isotropic, incompressible, elastic material. We will assume that the thin sheet deforms according to membrane theory. One of the basic assumptions for the membrane theory is that the material fibers which are orthogonal to the sheet deform in such a way that they are always orthogonal to the sheet. In addition, the stress components in the direction of the normal are much smaller in magnitude than the stress components in the tangential directions. We consider both these aspects next. In membrane theory the deformation of the sheet under pressure can be described by quantities which are related to the tangential directions to the mid-surface. The region

of the undeformed body of the flat sheet can be represented by

$$\mathbb{B} = \left\{ (x_1, x_2, x_3) : (x_1, x_2) \in \Omega, \quad |x_3| < \frac{h_0}{2} \right\}$$

with $x_3 = 0$, $\underline{x} = (x_1, x_2) \in \Omega$ being the mid-surface of the membrane. The mid-surface will deform as

$$(x_1, x_2, 0) \to (x_1 + u_1, x_2 + u_2, u_3) =: \underline{w}$$

where $u_i(\underline{x})$ for $i = 1, 2, 3$ denotes the **displacement** values, $\underline{w}$ denotes the **deformed mid-surface**. In the membrane model we only consider the mid-surface and the theory gives a 2D model of a 3D deformation.

The assumption about the stress in the membrane theory is that

$$\boldsymbol{\sigma}\underline{n} = \underline{0},$$

where $\underline{n}$ =**normal direction to the deformed state of the membrane**, and $\boldsymbol{\sigma}$ = **the membrane Cauchy stress**. Further details about the assumptions about fibres normal to the sheet remain normal to the mid-surface and the assumption that $\boldsymbol{\sigma}\underline{n} = \underline{0}$ are considered in the later sections. To be specific here, $\boldsymbol{\sigma}$ refers to the membrane stress and it relates to what we refer to as $\boldsymbol{\sigma}_{3D}$ evaluated on the mid-surface in a full three dimensional description. In a full three dimensional description without any simplifying assumptions the stress $\boldsymbol{\sigma}_{3D}$ varies through the thickness and satisfies traction boundary conditions on the lower and upper sides of the sheet. As already mentioned, the membrane model gives us a 2D model of the 3D deformation and in this context the pressure loading has a role of a "body force type term" in the membrane context and this should be more evident towards the end of section 3.4.2 when the weak form in the membrane case is first given.

The membrane deformation is described by the displacement $\underline{u}(\underline{x})$, $\underline{x} \in \Omega$ and the equations that determine the displacement field for a given loading in the case of a quasi-static deformation are the equations of equilibrium and the hyperelastic constitutive model for an incompressible material. As we will see, the material being incompressible is easily handled when we have a membrane deformation as the membrane stress $\boldsymbol{\sigma}$ is actually determined by the mid-surface displacement in order to be consistent with $\boldsymbol{\sigma}\underline{n} = \underline{0}$. A weak form of these equations is needed in this chapter and an outline of how these are obtained is as follows.

For a general 3D body **the equations of equilibrium** in the absence of body forces

are given by

$$\sum_{j=1}^{3} \frac{\partial \sigma_{ji}}{\partial w_j} = 0 \quad i = 1, 2, 3 \quad \text{or in vector form as} \quad \nabla \cdot \boldsymbol{\sigma}_{3D} = \underline{0}. \tag{3.2.1}$$

Note that the partial derivatives are with respect to the deformed configuration. To relate everything to the known undeformed configuration leads to the use of the first Piola stress which we explain later. A weak form of the system of equations can be obtained by taking the scalar product with a test vector $\underline{v}$ from an appropriate space and integrating over the domain. This is the general idea in all cases that we consider. When the undeformed thickness of the sheet, denoted by h_0, and the membrane assumptions for an incompressible body are taken into account we show later that the following is obtained.

For a given pressure $P(t)$ at a time t let $\mathbf{\Pi}$ denote the membrane nominal stress and let $\mathbf{\Pi}^T$ denote the membrane first Piola stress, which depends on the the displacement $\underline{u}$, and let the test space be given by

$$V = \{\underline{v} : \quad v_i \in H_0^1(\Omega), \quad i = 1, 2, 3\}, \tag{3.2.2}$$

where

$$H_0^1(\Omega) = \{\underline{v} : \underline{v} \in H^1(\Omega) \quad \text{and} \quad \underline{v} = \underline{0} \text{ on } \partial\Omega\}.$$

Also let

$$a_1(\underline{u}; \underline{v}) = h_0 \int_\Omega \mathbf{\Pi}^T : \nabla \underline{v} \, d\Omega, \quad a_2(\underline{u}; \underline{v}) = \int_\Omega \underline{v} \cdot \left(\frac{\partial \underline{w}}{\partial x_1} \times \frac{\partial \underline{w}}{\partial x_2} \right) d\Omega. \tag{3.2.3}$$

The notation $a_1(\cdot;\cdot)$ and $a_2(\cdot;\cdot)$ indicates that each of these is a semi-linear form which means that it is linear for the term to the right of the semi-colon, i.e. the expressions are nonlinear in $\underline{u}$ but linear in $\underline{v}$. For fixed time t, we then consider the following weak problem.

Find the displacement $\underline{u} \in V$ of the deformed membrane mid-surface, such that

$$A(t)(\underline{u}; \underline{v}) = a_1(\underline{u}; \underline{v}) - P(t) a_2(\underline{u}; \underline{v}) = \underline{0} \quad \forall \underline{v} \ \in V. \tag{3.2.4}$$

3.3 Background theory for a 3D solid

3.3.1 Membrane deformation and strain tensors

In this section we introduce some standard quantities used to describe the large deformation of a general body. More theory and material can be found in various sources such as [14], [25], [26], [12], [29] and many others.
We start by considering how line, surface and volume elements deform and in this part of the description we use the notation $\underline{x} \equiv (x_1, x_2, x_3)$ for a general point.

In the deformation, a point $\underline{x} \in \mathbb{B}$ moves to $\underline{x} + \underline{u}(\underline{x})$, where $\underline{u} \equiv (u_1, u_2, u_3)^T$ is the **displacement vector**. Therefore we have the following mapping

$$\underline{x} \longrightarrow \underline{x} + \underline{u}(\underline{x}) \equiv \underline{w}, \tag{3.3.1}$$

where $\underline{w}$ denotes the deformed position.

The **deformation gradient in 3D** involves all the first partial derivatives of the components of $\underline{u}$ and is defined by

$$\mathbf{F}_{3D} = \mathbf{I} + \nabla \underline{u}, \tag{3.3.2}$$

where $\mathbf{I}$ is the identity tensor. In cartesian components $\mathbf{I}$ and $\nabla \underline{u}$ are given by

$$\mathbf{I} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad \nabla \underline{u} = \begin{pmatrix} \frac{\partial u_1}{\partial x_1} & \frac{\partial u_1}{\partial x_2} & \frac{\partial u_1}{\partial x_3} \\ \frac{\partial u_2}{\partial x_1} & \frac{\partial u_2}{\partial x_2} & \frac{\partial u_2}{\partial x_3} \\ \frac{\partial u_3}{\partial x_1} & \frac{\partial u_3}{\partial x_2} & \frac{\partial u_3}{\partial x_3} \end{pmatrix}. \tag{3.3.3}$$

Here, $\mathbf{F}_{3D}$ corresponds to the Jacobian matrix of the mapping (3.3.1). It is an important quantity for measuring large deformation. By using $\mathbf{F}_{3D}$ we are able to define the 3D **Right Cauchy Green deformation tensor** given by

$$\mathbf{C}_{3D} = \mathbf{F}_{3D}^T \mathbf{F}_{3D}. \tag{3.3.4}$$

Finally we have the **Left Cauchy Green deformation tensor** which is denoted by

$$\mathbf{B}_{3D} = \mathbf{F}_{3D} \mathbf{F}_{3D}^T, \tag{3.3.5}$$

from which it follows that

$$\mathbf{B}_{3D}^{-1} = (\mathbf{F}_{3D}^{-1})^T \mathbf{F}_{3D}^{-1}. \tag{3.3.6}$$

We use the notation $\mathbf{F}_{3D}$, $\mathbf{C}_{3D}$ and $\mathbf{B}_{3D}$ here as we use the notation $\mathbf{F}$, $\mathbf{C}$ for corresponding quantities in the membrane description. Now, by using the mapping (3.3.1) we have in terms of infinitesimals the following representation

$$\Delta\underline{w} = \underline{w}(\underline{x} + \Delta\underline{x}) - \underline{w}(\underline{x}) = \mathbf{F}_{3D}\Delta\underline{x} \tag{3.3.7}$$

The above represents the infinitesimal line segment $\Delta\underline{x}$ which deforms to the line segment $\Delta\underline{w} = \mathbf{F}_{3D}\Delta\underline{x}$.

Let $\underline{e}_1, \underline{e}_2, \underline{e}_3$ denote the unit base vectors with respect to the x_1, x_2, x_3 directions. We start by considering the particular line segments $\Delta\underline{x}_1 = \Delta x_1\underline{e}_1$, $\Delta\underline{x}_2 = \Delta x_2\underline{e}_2$, $\Delta\underline{x}_3 = \Delta x_3\underline{e}_3$ having directions coinciding with the coordinate directions. In this case the surface element $\Delta\underline{x}_1 \times \Delta\underline{x}_2$ and the volume element $(\Delta\underline{x}_1 \times \Delta\underline{x}_2) \cdot \Delta\underline{x}_3$ transform respectively to $\Delta\underline{w}_1 \times \Delta\underline{w}_2$ and $(\Delta\underline{w}_1 \times \Delta\underline{w}_2) \cdot \Delta\underline{w}_3$ as follows.

$$\begin{aligned}
\Delta\underline{w}_1 \times \Delta\underline{w}_2 &= \Delta x_1\Delta x_2\mathbf{F}_{3D}\underline{e}_1 \times \mathbf{F}_{3D}\underline{e}_2 && (3.3.8)\\
&= \Delta x_1\Delta x_2(\det\mathbf{F}_{3D})\mathbf{F}_{3D}{}^{-T}\underline{e}_1 \times \underline{e}_2 && (3.3.9)\\
&= (\det\mathbf{F}_{3D})\mathbf{F}_{3D}{}^{-T}\Delta\underline{x}_1 \times \Delta\underline{x}_2 && (3.3.10)
\end{aligned}$$

and

$$\begin{aligned}
(\Delta\underline{w}_1 \times \Delta\underline{w}_2) \cdot \Delta\underline{w}_3 &= (\det\mathbf{F}_{3D})(\Delta\underline{x}_1 \times \Delta\underline{x}_2)\mathbf{F}_{3D}{}^{-1}\mathbf{F}_{3D}\Delta\underline{x}_3 && (3.3.11)\\
&= (\det\mathbf{F}_{3D})(\Delta\underline{x}_1 \times \Delta\underline{x}_2) \cdot \Delta\underline{x}_3. && (3.3.12)
\end{aligned}$$

Note that here we used the result that

$$\mathbf{F}_{3D}\underline{e}_1 \times \mathbf{F}_{3D}\underline{e}_2 = (\det\mathbf{F}_{3D})\mathbf{F}_{3D}{}^{-T}\underline{e}_1 \times \underline{e}_2. \tag{3.3.13}$$

Proof of eq.(3.3.13)

$$(\det\mathbf{F}_{3D})\mathbf{F}_{3D}{}^{-T} = \text{the matrix of cofactors}$$

Since $\underline{e}_1 \times \underline{e}_2 = \underline{e}_3$,

$$(\det\mathbf{F}_{3D})\mathbf{F}_{3D}{}^{-T}\underline{e}_1 \times \underline{e}_2 = (\det\mathbf{F}_{3D})\mathbf{F}_{3D}{}^{-T}\underline{e}_3 = \begin{pmatrix} F_{21}F_{32} - F_{31}F_{22} \\ F_{12}F_{31} - F_{32}F_{11} \\ F_{11}F_{22} - F_{12}F_{21} \end{pmatrix}. \tag{3.3.14}$$

By using the cross product $\mathbf{F}_{3D}\underline{e}_1 \times \mathbf{F}_{3D}\underline{e}_2$ we get

$$\mathbf{F}_{3D}\underline{e}_1 \times \mathbf{F}_{3D}\underline{e}_2 = \begin{vmatrix} \underline{e}_1 & \underline{e}_2 & \underline{e}_3 \\ F_{11} & F_{21} & F_{31} \\ F_{21} & F_{22} & F_{32} \end{vmatrix} = \begin{pmatrix} F_{21}F_{32} - F_{31}F_{22} \\ F_{12}F_{31} - F_{32}F_{11} \\ F_{11}F_{22} - F_{12}F_{21} \end{pmatrix}. \tag{3.3.15}$$

Combining eq.(3.3.14) and (3.3.15) we get the desired result eq.(3.3.13).

The above relations for the surface and volume elements can be generalised and given in terms of line segments in arbitrary directions. More precisely, we get the following relationships between the undeformed and deformed line segments $\Delta\underline{x}$ and $\Delta\underline{w}$, the undeformed and deformed surface elements $\underline{N}\Delta S$ and $\underline{n}\Delta s$ and the undeformed and deformed volume elements ΔV and Δv.

$$\Delta\underline{w} = \mathbf{F}_{3D}\Delta\underline{x} \tag{3.3.16}$$

$$\underline{n}\Delta s = (\det\mathbf{F}_{3D})\mathbf{F}_{3D}^{-T}\underline{N}\Delta S \tag{3.3.17}$$

$$\Delta v = (\det\mathbf{F}_{3D})\Delta V. \tag{3.3.18}$$

3.3.2 Decomposition of a deformation and principal axes

By using the polar decomposition theorem, the deformation tensor $\mathbf{F}_{3D}$ can be expressed in the following polar forms

$$\mathbf{F}_{3D} = \mathbf{R}\cdot\mathbf{U} = \mathbf{V}\cdot\mathbf{R} \tag{3.3.19}$$

where $\mathbf{R}$ is an orthogonal tensor and $\mathbf{U}, \mathbf{V}$ are both symmetric positive definite tensors. If ρ_0 denotes the density in the undeformed configuration and ρ denotes the density in the deformed configuration then the conservation of mass and the relation (3.3.18) implies that $\det\mathbf{F}_{3D} = \rho_0/\rho > 0$, and therefore $\mathbf{F}_{3D}$ is a proper orthogonal tensor.

By using the polar forms and the definitions of $\mathbf{C}_{3D}$ and $\mathbf{B}_{3D}$ we get the following relation

$$\mathbf{C}_{3D} = \mathbf{F}_{3D}^T\mathbf{F}_{3D} = (\mathbf{RU})^T\mathbf{RU} = \mathbf{U}^T\mathbf{R}^T\mathbf{RU} = \mathbf{U}^T\mathbf{U} = \mathbf{U}^2 \tag{3.3.20}$$

and

$$\mathbf{B}_{3D} = \mathbf{F}_{3D}\mathbf{F}_{3D}^T = \mathbf{VR}(\mathbf{VR})^T = \mathbf{VRR}^T\mathbf{V}^T = \mathbf{VV}^T = \mathbf{V}^2 \tag{3.3.21}$$

where we used $\mathbf{R}^T\mathbf{R} = \mathbf{RR}^T = \mathbf{I}$ since $\mathbf{R}$ is orthogonal matrix. Also we can define the following relations for $\mathbf{U}$ and $\mathbf{V}$

$$\mathbf{U} = \mathbf{R}^T\mathbf{F}_{3D} = \mathbf{R}^T\mathbf{VR} \quad \text{and} \quad \mathbf{V} = \mathbf{F}_{3D}\mathbf{R}^T = \mathbf{RUR}^T. \tag{3.3.22}$$

That is $\mathbf{U}$ and $\mathbf{V}$ are similar and therefore they have the same positive eigenvalues $\lambda_1, \lambda_2, \lambda_3$ known as the **principal stretches**. We let $\underline{v}_1, \underline{v}_2, \underline{v}_3$ be eigenvectors of $\mathbf{U}$ of unit length with the corresponding eigenvalues $\lambda_1, \lambda_2, \lambda_3$. It follows that the eigenvectors of $\mathbf{V}$ are $\hat{\underline{v}}_1 = \mathbf{R}\underline{v}_1$, $\hat{\underline{v}}_2 = \mathbf{R}\underline{v}_2$, $\hat{\underline{v}}_3 = \mathbf{R}\underline{v}_3$. The eigenvectors of $\mathbf{U}$ and $\mathbf{V}$ give the **principal directions of stretch** in the undeformed and deformed configurations respectively.

Since $\mathbf{C}_{3D} = \mathbf{U}^2$, the eigenvectors of $\mathbf{C}_{3D}$ coincide with those of $\mathbf{U}$ and its eigenvalues are just $\lambda_1^2, \lambda_2^2, \lambda_3^2$. Similarly, since $\mathbf{B}_{3D} = \mathbf{V}^2$, the eigenvectors of $\mathbf{B}_{3D}$ coincide with those of $\mathbf{V}$ and its eigenvalues are just $\lambda_1^2, \lambda_2^2, \lambda_3^2$. Using these eigenvalues and eigenvectors, we have the following spectral decompositions and singular-valued decompositions:

$$\mathbf{F}_{3D} = \lambda_1 \hat{\underline{v}}_1 \underline{v}_1^T + \lambda_2 \hat{\underline{v}}_2 \underline{v}_2^T + \lambda_3 \hat{\underline{v}}_3 \underline{v}_3^T \tag{3.3.23}$$

$$\mathbf{C}_{3D} = \lambda_1^2 \underline{v}_1 \underline{v}_1^T + \lambda_2^2 \underline{v}_2 \underline{v}_2^T + \lambda_3^2 \underline{v}_3 \underline{v}_3^T \tag{3.3.24}$$

$$\mathbf{U} = \lambda_1 \underline{v}_1 \underline{v}_1^T + \lambda_2 \underline{v}_2 \underline{v}_2^T + \lambda_3 \underline{v}_3 \underline{v}_3^T \tag{3.3.25}$$

$$\mathbf{B}_{3D} = \lambda_1^2 \hat{\underline{v}}_1 \hat{\underline{v}}_1^T + \lambda_2^2 \hat{\underline{v}}_2 \hat{\underline{v}}_2^T + \lambda_3^2 \hat{\underline{v}}_3 \hat{\underline{v}}_3^T \tag{3.3.26}$$

$$\mathbf{V} = \lambda_1 \hat{\underline{v}}_1 \hat{\underline{v}}_1^T + \lambda_2 \hat{\underline{v}}_2 \hat{\underline{v}}_2^T + \lambda_3 \hat{\underline{v}}_3 \hat{\underline{v}}_3^T \tag{3.3.27}$$

Remark: In the case of an incompressible deformation we have

$$\det \mathbf{F}_{3D} = \lambda_1 \lambda_2 \lambda_3 = 1. \tag{3.3.28}$$

Strain Invariants

Here, we define the **principal strain invariants** I_1, I_2, I_3 which are related to the 3 principal stretches $\lambda_1, \lambda_2, \lambda_3$ and are given by

$$I_1 = \lambda_1^2 + \lambda_2^2 + \lambda_3^2 \tag{3.3.29}$$

$$I_2 = \lambda_1^2 \lambda_2^2 + \lambda_1^2 \lambda_3^2 + \lambda_2^2 \lambda_3^2 \tag{3.3.30}$$

$$I_3 = \lambda_1^2 \lambda_2^2 \lambda_3^2. \tag{3.3.31}$$

Since λ_1^2, λ_2^2 and λ_3^2 are the principal values of both $\mathbf{C}_{3D}$ and $\mathbf{B}_{3D}$ we have the following

alternative expressions.

$$I_1 = \mathrm{tr}(\mathbf{C}_{3D}) = \mathrm{tr}(\mathbf{B}_{3D}) \tag{3.3.32}$$

$$I_2 = \frac{1}{2}\left((\mathrm{tr}(\mathbf{C}_{3D}^2) - \mathrm{tr}(\mathbf{C}_{3D}^2)\right) = \frac{1}{2}\left((\mathrm{tr}(\mathbf{B}_{3D}^2) - \mathrm{tr}(\mathbf{B}_{3D}^2)\right) \tag{3.3.33}$$

$$I_3 = \det(\mathbf{C}_{3D}) = \det(\mathbf{B}_{3D}) \tag{3.3.34}$$

In addition, since

$$I_3 = \det(\mathbf{C}_{3D}) = \det(\mathbf{F}_{3D}^T\mathbf{F}_{3D}) = (\det\mathbf{F}_{3D})^2 = \left(\frac{dv}{dV}\right)^2 = \left(\frac{\rho_0}{\rho}\right)^2,$$

we have for an incompressible material that $I_3 = 1$. We do not actually make use of I_1, I_2 and I_3 much in this thesis as we always work with λ_1, λ_2 and λ_3 and use a hyperelastic model in an Ogden form but some of the hyperelastic models that are mentioned involve expressions in I_1 and I_2. We note here that, assuming incompressibility, we have the following relation between the principal values

$$\lambda_3 = 1/\left(\lambda_1\lambda_2\right).$$

3.3.3 Stress tensors

First we start with the definition of a stress, which is the force acting on an interior point, of a continuous body $\mathbb{B}$ from its neighbouring parts. We consider a surface element area Δs with unit normal $\underline{n}$, on which the material outside exerts a force $\Delta\mathbf{f}$ such that

$$\Delta\mathbf{f} = \underline{\tau}^n\Delta s, \tag{3.3.35}$$

where $\underline{\tau}^n$ denotes the mean surface traction across the element of area Δs. By taking the limit as $\Delta s \to 0$, it is assumed that $\underline{\tau}^n$ tends to a finite limit. The traction vector, at the point, on the surface Δs with normal $\underline{n}$ is given by

$$\underline{\tau}^n = \lim_{\Delta s\to 0}\frac{\Delta\mathbf{f}}{\Delta s}. \tag{3.3.36}$$

An infinite number of traction vectors act at a point with each acting on different surfaces through the point, defined by different normals.

Cauchy's Law states that there exists a Cauchy stress tensor $\boldsymbol{\sigma}_{3D}$ which maps the

normal vector to a surface to the traction vector acting on that surface, i.e. we have

$$\underline{\tau} = \boldsymbol{\sigma}_{3D}\underline{n} \quad \text{or} \quad \tau_i = \sigma_{ij}n_j, \tag{3.3.37}$$

assuming the summation connection. In full we have

$$\tau_1 = \sigma_{11}n_1 + \sigma_{12}n_2 + \sigma_{13}n_3 \tag{3.3.38}$$

$$\tau_2 = \sigma_{21}n_1 + \sigma_{22}n_2 + \sigma_{23}n_3 \tag{3.3.39}$$

$$\tau_3 = \sigma_{31}n_1 + \sigma_{32}n_2 + \sigma_{33}n_3 \tag{3.3.40}$$

The components of the stress tensor $\boldsymbol{\sigma}_{3D}$ with respect to a Cartesian coordinate system are

$$\sigma_{ij} = \underline{e}_i \cdot \boldsymbol{\sigma}_{3D}\underline{e}_j = \underline{e}_i \cdot \underline{\tau}^{\underline{e}_j} \tag{3.3.41}$$

which is the i^{th} component of the traction vector acting on a surface with normal $\underline{e}_j$.

The three traction vectors acting on the surface elements whose outward normals point in the directions of the three base vectors $\underline{e}_j$ are

$$\underline{\tau}^{\underline{e}_j} = \boldsymbol{\sigma}_{3D}\underline{e}_j, \quad j = 1, 2, 3 \tag{3.3.42}$$

or in full

$$\underline{\tau}^{\underline{e}_1} = \sigma_{11}\underline{e}_1 + \sigma_{21}\underline{e}_2 + \sigma_{31}\underline{e}_3 \tag{3.3.43}$$

$$\underline{\tau}^{\underline{e}_2} = \sigma_{12}\underline{e}_1 + \sigma_{22}\underline{e}_2 + \sigma_{32}\underline{e}_3 \tag{3.3.44}$$

$$\underline{\tau}^{\underline{e}_3} = \sigma_{13}\underline{e}_1 + \sigma_{23}\underline{e}_2 + \sigma_{33}\underline{e}_3 \tag{3.3.45}$$

The proof of the Cauchy's Law and more details about the stress components can be found in [29]. Although the Cauchy stress tensor is defined in the deformed configuration, in our membrane model we used stress tensors which are defined in the undeformed configuration. By using Cauchy's Law, a surface element $\underline{n}\Delta s$ in the deformed configuration has a force on it given by $\boldsymbol{\sigma}_{3D}\underline{n}\Delta s$. Using eq. (3.3.17) we have that

$$\boldsymbol{\sigma}_{3D}\underline{n}\Delta s = \boldsymbol{\sigma}_{3D}(\det\mathbf{F}_{3D})\mathbf{F}_{3D}^{-T}\underline{N}\Delta S = \boldsymbol{\Pi}_{3D}^{T}\underline{N}\Delta S \tag{3.3.46}$$

where

$$\boldsymbol{\Pi}_{3D}^{T} = \boldsymbol{\sigma}_{3D}(\det\mathbf{F}_{3D})\mathbf{F}_{3D}^{-T} = \textbf{The First Piola stress tensor}. \tag{3.3.47}$$

If we take the transpose of the first Piola stress tensor we get the following stress tensor

$$\mathbf{\Pi}_{3D} = (\det\mathbf{F}_{3D})\mathbf{F}_{3D}^{-1}\boldsymbol{\sigma}_{3D} = \textbf{Nominal stress tensor}. \tag{3.3.48}$$

3.3.4 The equations of motion and the weak form

By applying the Cauchy's Law $\underline{\tau} = \boldsymbol{\sigma}_{3D}\underline{n}$ and the divergence theorem (2.5.1) we get the global form of **the equations of motion**,

$$\frac{\partial\sigma_{ij}}{\partial\omega_j} + b_i = \rho\frac{d\dot{u}_i}{dt} \quad \text{or} \quad \nabla\cdot\boldsymbol{\sigma}_{3D} + \underline{b} = \rho\frac{d\underline{\dot{u}}}{dt}, \tag{3.3.49}$$

where $\underline{b}$ represents the body forces and $\frac{d\underline{\dot{u}}}{dt}$ is the acceleration at a point.

Now for the special case when the acceleration is zero, the equations reduce to **the equations of equilibrium** which have the following form

$$\frac{\partial\sigma_{ij}}{\partial\omega_j} + b_i = 0 \quad \text{or} \quad \nabla\cdot\boldsymbol{\sigma}_{3D} + \underline{b} = \underline{0}. \tag{3.3.50}$$

In our applications we consider the case of the quasi-static deformation with no body forces ($\underline{b} = \underline{0}$).

The equilibrium equations involve the partial derivatives of the Cauchy stress with respect to the components w_j of the deformed position. In order to convert this relation to quantities which relate to the undeformed configuration we do the following.

For the case of no body forces ($\underline{b} = \underline{0}$) and an incompressible deformation ($\det\mathbf{F}_{3D} = 1$) we have

$$\nabla\cdot\boldsymbol{\sigma}_{3D} = \underline{0} \tag{3.3.51}$$

As this holds at every point in the domain this implies that for any volume v with surface s we have

$$\int_s \boldsymbol{\sigma}_{3D}^T\underline{n}\,ds = \underline{0}. \tag{3.3.52}$$

Then, by using the identity (3.3.46), which defines the relationship between the surface elements in the deformed and the undeformed configuration we get

$$\int_S \mathbf{\Pi}_{3D}^T\underline{N}\,dS = \underline{0}. \tag{3.3.53}$$

As the regions involved are arbitrary, the equilibrium equations can be written as

$$\operatorname{Div}\mathbf{\Pi}_{3D} = \begin{pmatrix} \frac{\partial \Pi_{11}}{\partial x_1} + \frac{\partial \Pi_{21}}{\partial x_2} + \frac{\partial \Pi_{31}}{\partial x_3} \\ \frac{\partial \Pi_{12}}{\partial x_1} + \frac{\partial \Pi_{22}}{\partial x_2} + \frac{\partial \Pi_{32}}{\partial x_3} \\ \frac{\partial \Pi_{13}}{\partial x_1} + \frac{\partial \Pi_{23}}{\partial x_2} + \frac{\partial \Pi_{33}}{\partial x_3} \end{pmatrix} = \underline{0}. \tag{3.3.54}$$

Weak formulation for the 3D case

In order to obtain the weak form of the eq. (3.3.54), we do the following.
First we define the following function space

$$V = \left\{ \underline{v} : v_i \in H_0^1(\Omega_{3D}),\, v_i(\underline{x}) = \underline{0} \text{ on } \partial\mathbb{B} \quad i = 1, 2, 3 \right\}.$$

Then, by taking the dot product with a test vector $\underline{v} \in V$, integrating over the region, using the divergence theorem (2.5.1) and the following identity

$$\operatorname{Div}(\mathbf{\Pi}_{3D}\underline{v}) = \operatorname{Div}\mathbf{\Pi}_{3D} \cdot \underline{v} + \mathbf{\Pi}_{3D}^T : \nabla\underline{v}, \quad \forall\, \underline{v} \in V,$$

we end up with the following weak form in 3D

$$\int_{\mathbb{B}} \mathbf{\Pi}_{3D}^T : \nabla\underline{v}\, dV - \int_{\partial\mathbb{B}} \underline{v}^T \mathbf{\Pi}_{3D}^T \underline{N} dS = 0 \quad \forall \underline{v} \in V, \tag{3.3.55}$$

where $\mathbb{B}$ represents the region of the undeformed body and where : denotes the double dot product operation to get a scalar by combining $\mathbf{\Pi}_{3D}^T$ and $\nabla\underline{v}$. Note that when $\underline{u}$ is given on the boundary $\partial\mathbb{B}$, the space V is such that $\underline{v} = \underline{0}$ on $\partial\Omega$ and thus the boundary integral term vanishes.

3.4 Simplifications for the membrane model

In the previous sections we described the general case of the deformation of an arbitrary 3D body. However, for this project we only need to consider how thin sheets deform under pressure loading. According to the membrane theory, the material fibres which are normal to the mid-surface Ω in the undeformed state and remain normal to whatever the deformed mid-surface is, but they will usually change in length by the factor $\lambda = \lambda_3$ which is known as **the thickness stretch ratio**. When we have incompressibility λ_3 is determined by the stretch ratios in directions tangential to the mid-surface. This suggests that the deformation of the sheet can essentially be described in terms of quantities which just relate to the tangential directions which is what is done in membrane theory. With

$\boldsymbol{\sigma}_{3D}$ being the stress in the 3D model evaluated on the mid-surface and with $\boldsymbol{\sigma}$ being the membrane stress, the basic assumption is that

$$\boldsymbol{\sigma}_{3D}\underline{n} \approx \underline{0},$$

with $\underline{n}$ being the unit outward normal to the mid-surface. Approximately $\underline{0}$ is in the sense that it is much smaller in magnitude than $\boldsymbol{\sigma}_{3D}\underline{d}$ for any unit vector $\underline{d}$ tangential to the mid-surface. In the membrane idealization the membrane stress $\boldsymbol{\sigma}$ is such that

$$\boldsymbol{\sigma}\underline{n} = \underline{0}.$$

It is this assumed behaviour concerning how the sheet deforms in the normal direction which leads from a 3D problem to a 2D problem and a 3D axisymmetric problem to be reduced to a 1D problem. The axisymmetric case of the unconstrained inflation of membranes is considered later in the thesis. More details about the membrane theory can be found in [13].

3.4.1 Membrane quantities and weak formulation in 2D

The **membrane deformation gradient** in 2D evaluated on the mid-surface is given by

$$\mathbf{F} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} + \nabla \underline{u} = \begin{pmatrix} 1 + \frac{\partial u_1}{\partial x_1} & \frac{\partial u_1}{\partial x_2} \\ \frac{\partial u_2}{\partial x_1} & 1 + \frac{\partial u_2}{\partial x_2} \\ \frac{\partial u_3}{\partial x_1} & \frac{\partial u_3}{\partial x_2} \end{pmatrix}, \tag{3.4.1}$$

where $\mathbf{F}\underline{e}_1$ and $\mathbf{F}\underline{e}_2$ represent the tangential vectors of the model.
The membrane **Right Cauchy Green deformation tensor** is given by

$$\mathbf{C} = \mathbf{F}^T\mathbf{F} \tag{3.4.2}$$

and we let λ_1^2, λ_2^2 denote the eigenvalues of $\mathbf{C}$.

To describe instead the membrane deformation in the terms used to describe a general 3D deformation can be done as follows. The thickness stretch ratio

$$\lambda = \lambda_3 = 1/(\lambda_1\lambda_2),$$

and the 3D version of the deformation gradient is given by

$$\mathbf{F}_{3D} = (\mathbf{F}, \lambda \underline{n}) = \begin{pmatrix} 1 + \frac{\partial u_1}{\partial x_1} & \frac{\partial u_1}{\partial x_2} & \lambda n_1 \\ \frac{\partial u_2}{\partial x_1} & 1 + \frac{\partial u_2}{\partial x_2} & \lambda n_2 \\ \frac{\partial u_3}{\partial x_1} & \frac{\partial u_3}{\partial x_2} & \lambda n_3 \end{pmatrix} \tag{3.4.3}$$

with n_1, n_2 and n_3 being the components of the unit outward normal vector $\underline{n}$. The vector $\underline{n}$ is orthogonal to $\mathbf{F}\underline{e}_1$ and $\mathbf{F}\underline{e}_2$ and the connection between these vectors is that

$$\underline{n} = \lambda \mathbf{F}\underline{e}_1 \times \mathbf{F}\underline{e}_2.$$

By using $\mathbf{F}_{3D}$ we are able to define the 3D Right Cauchy Green deformation tensor given by

$$\mathbf{C}_{3D} = \mathbf{F}_{3D}^T \mathbf{F}_{3D} = \begin{pmatrix} c_{11} & c_{21} & 0 \\ c_{21} & c_{22} & 0 \\ 0 & 0 & \lambda^2 \end{pmatrix}. \tag{3.4.4}$$

For an incompressible material $\det \mathbf{C}_{3D} = \lambda^2(c_{11}c_{22} - c_{21}^2) = 1$, which gives λ in terms of other components of $\mathbf{C}_{3D}$. The eigenvalues of $\mathbf{C}_{3D}$ are λ_1^2, λ_2^2 and λ^2.

A few other relations in a 3D description worth mentioning here quickly follow. Firstly, since $\mathbf{C}_{3D} = \mathbf{U}^2$, $\mathbf{U}$ is of the form

$$\mathbf{U} = \begin{pmatrix} U_{11} & U_{21} & 0 \\ U_{21} & U_{22} & 0 \\ 0 & 0 & \lambda \end{pmatrix}. \tag{3.4.5}$$

In addition from the polar decomposition of $\mathbf{F}_{3D}$, we have

$$\mathbf{F}_{3D}\underline{e}_3 = \mathbf{R}\mathbf{U}\underline{e}_3 = \lambda \mathbf{R}\underline{e}_3,$$

which implies that the third column of $\mathbf{R}$ is the unit vector in the normal direction to the mid-surface Ω, i.e. $\underline{n} = R\underline{e}_3$. If we let $\underline{v}_1$ and $\underline{v}_2$ be normalised eigenvectors of C_{3D} corresponding to eigenvalues λ_1^2 and λ_2^2 respectively and we let $\hat{\underline{v}}_1 = R\underline{v}_1$ and $\hat{\underline{v}}_2 = R\underline{v}_2$ then the spectral decompositions and singular valued decompositions mentioned in (3.3.23) and (3.3.24) are given by

$$\mathbf{F}_{3D} = \lambda_1 \hat{\underline{v}}_1 \, \underline{v}_1^T + \lambda_2 \hat{\underline{v}}_2 \, \underline{v}_2^T + \lambda \underline{n} \, \underline{e}_3^T, \tag{3.4.6}$$

$$\mathbf{C}_{3D} = \lambda_1^2 \underline{v}_1 \, \underline{v}_1^T + \lambda_2^2 \underline{v}_2 \, \underline{v}_2^T + \lambda^2 \underline{e}_3 \, \underline{e}_3^T. \tag{3.4.7}$$

and it is worth noting here that we also have

$$\mathbf{F}_{3D}^{-T} = \frac{1}{\lambda_1}\hat{\underline{v}}_1\,\underline{v}_1^T + \frac{1}{\lambda_2}\hat{\underline{v}}_2\,\underline{v}_2^T + \frac{1}{\lambda}\underline{n}\,\underline{e}_3^T \tag{3.4.8}$$

and in particular we have that $\underline{n}$ is given via

$$\mathbf{F}_{3D}^{-T}\underline{e}_3 = \frac{1}{\lambda}\underline{n}. \tag{3.4.9}$$

In the case of the stress, in the full 3D model we have $\boldsymbol{\sigma}_{3D}\underline{n} \approx \underline{0}$, which implies that

$$\boldsymbol{\sigma}_{3D}\underline{n} = \boldsymbol{\sigma}_{3D}\mathbf{R}\underline{e}_3 \approx \underline{0}$$

and the membrane version of this is that $\boldsymbol{\sigma}\underline{n} = \underline{0}$, which implies that

$$\boldsymbol{\sigma}\underline{n} = \boldsymbol{\sigma}\mathbf{R}\underline{e}_3 = \underline{0},$$

and therefore for the symmetric tensor $\mathbf{R}^T\boldsymbol{\sigma}\mathbf{R}$ we have that

$$\mathbf{R}^T\boldsymbol{\sigma}\mathbf{R} = \begin{pmatrix} \widetilde{\sigma_{11}} & \widetilde{\sigma_{21}} & 0 \\ \widetilde{\sigma_{21}} & \widetilde{\sigma_{22}} & 0 \\ 0 & 0 & 0 \end{pmatrix}, \tag{3.4.10}$$

which involves only 3 independent stress components. Further, it is worth noting that the condition $\boldsymbol{\sigma}_{3D}\underline{n} \approx \underline{0}$ on the Cauchy stress corresponds to the condition

$$\boldsymbol{\Pi}_{3D}^T\underline{e}_3 \approx \underline{0} \tag{3.4.11}$$

on the 3D version of the first Piola stress for the undeformed configuration of the membrane.

3.4.2 Membrane deformation under pressure loading

We consider now the quasi-static equilibrium when there is an applied pressure P on the lower side of the sheet. On this side of the sheet we consider the following conditions

$$\underline{N}_{low} = -\underline{e}_3 \quad \text{and} \quad \underline{n}_{low} \approx -\underline{n},$$

which denote the normals to the undeformed and the deformed lower side of the sheet. In the 3D case the traction boundary condition is

$$\boldsymbol{\sigma}_{3D}^{T}\underline{n}_{low} = \boldsymbol{\sigma}_{3D}\underline{n}_{low} = -P\underline{n}_{low}.$$

Re-writing in terms of the undeformed coordinates gives

$$\boldsymbol{\Pi}_{3D}^{T}\underline{N}_{low} = \boldsymbol{\sigma}^{T}\underline{n}_{low}\frac{ds}{dS} = -P\underline{n}_{low}\frac{ds}{dS} \approx -P(-\underline{n})/\lambda = P\underline{n}/\lambda = P\mathbf{F}_{3D}^{-T}\underline{e}_3. \tag{3.4.12}$$

Next we define the reduced form of the weak form which in the full 3D case is given in (3.3.55). By using the relation (3.4.12) we are able to approximate the boundary part of the weak form by the following expression

$$\int_{\partial\mathbb{B}} \underline{v}^{T}\boldsymbol{\Pi}_{3D}^{T}\underline{N}\, dS \approx P\iint_{\Omega}\left(\underline{v}^{T}\mathbf{F}_{3D}^{-T}\underline{e}_3\right)\, dx_1dx_2, \tag{3.4.13}$$

which is what we call the **pressure term** of our model.

The term involving the region of the undeformed body $\mathbb{B}$ can be written as

$$\int_{\mathbb{B}} \boldsymbol{\Pi}_{3D}^{T} : \nabla\underline{v}\, dV \approx h_0\iint_{\Omega}\boldsymbol{\Pi}^{T} : \nabla\underline{v}\, dx_1dx_2 \tag{3.4.14}$$

which involves the thickness of the membrane and the mid-surface Ω. Here $\boldsymbol{\Pi}^{T}$ is 3×2 in shape and represents the **membrane first Piola Stress**.

Finally we are able to define the weak form of the membrane case which has the following form.
Find $\underline{u}\in V$ such that

$$A(t)(\underline{u};\underline{v}) = a_1(\underline{u};\underline{v}) - Pa_2(\underline{u};\underline{v}) = 0 \quad \forall\underline{v}\in V, \tag{3.4.15}$$

$$a_1(\underline{u};\underline{v}) = h_0\iint_{\Omega}\boldsymbol{\Pi}^{T} : \nabla\underline{v}\, dx_1dx_2 \quad \forall\underline{v}\in V, \tag{3.4.16}$$

$$a_2(\underline{u};\underline{v}) = \iint_{\Omega}\underline{v}\cdot\left(\frac{\partial\underline{w}}{\partial x_1}\times\frac{\partial\underline{w}}{\partial x_2}\right)\, dx_1dx_2 \quad \forall\underline{v}\in V, \tag{3.4.17}$$

with the test space V given by

$$V = \{\underline{v} : v_i\in H_0^1(\Omega), \quad v_i(x_1,x_2) = 0 \text{ when } (x_1,x_2)\in\partial\Omega \quad i=1,2,3\}.$$

Note, from the representation of the weak form, we observe that the membrane deformation depends only on the ratio P/h_0.

There are other ways the pressure loading term in the weak form can be written and we give these next. For the pressure term (3.4.17) we used the following relations. Possibly the simplest relation is to note that

$$\mathbf{F}_{3D}^{-T}\underline{e}_3 = \frac{1}{\lambda}\underline{n} = \mathbf{F}\underline{e}_1 \times \mathbf{F}\underline{e}_2 = \left(\frac{\partial \underline{w}}{\partial x_1} \times \frac{\partial \underline{w}}{\partial x_2}\right).$$

In the weak form we have the scalar product of this with the test vector $\underline{v}$ and for the integral of this over Ω we have the following, which we will use at several points of this thesis.

$$\begin{aligned}\iint_\Omega \underline{v} \cdot \left(\frac{\partial \underline{w}}{\partial x_1} \times \frac{\partial \underline{w}}{\partial x_2}\right) dx_1 dx_2 = \frac{1}{3}\iint_\Omega \underline{v} \cdot \left(\frac{\partial \underline{w}}{\partial x_1} \times \frac{\partial \underline{w}}{\partial x_2}\right) dx_1 dx_2 \\ +\frac{1}{3}\iint_\Omega \underline{w} \cdot \left(\frac{\partial \underline{v}}{\partial x_1} \times \frac{\partial \underline{w}}{\partial x_2}\right) dx_1 dx_2 + \frac{1}{3}\iint_\Omega \underline{w} \cdot \left(\frac{\partial \underline{w}}{\partial x_1} \times \frac{\partial \underline{v}}{\partial x_2}\right) dx_1 dx_2. \end{aligned} \tag{3.4.18}$$

Although this is a longer expression, it enables the expression in the weak form to be written in a more symmetrical way which we comment on later when some computational details are described.

Proof of the expression eq. (3.4.18)
We start with the following quantity

$$I = \iint_\Omega \underline{w} \cdot \left(\frac{\partial \underline{v}}{\partial x_1} \times \frac{\partial \underline{w}}{\partial x_2} + \frac{\partial \underline{w}}{\partial x_1} \times \frac{\partial \underline{v}}{\partial x_2}\right) dx_1 dx_2, \tag{3.4.19}$$

which is part of the second version and we show that

$$I = 2\iint_\Omega \underline{v} \cdot \left(\frac{\partial \underline{w}}{\partial x_1} \times \frac{\partial \underline{w}}{\partial x_2}\right) dx_1 dx_2 \tag{3.4.20}$$

to get the desired result.

By using an invariance properties of the scalar triple product when the terms are re-ordered, the integrand of the quantity eq.(3.4.19) can be written as

$$\underline{w} \cdot \left[\left(\frac{\partial \underline{v}}{\partial x_1} \times \frac{\partial \underline{w}}{\partial x_2}\right) + \left(\frac{\partial \underline{w}}{\partial x_1} \times \frac{\partial \underline{v}}{\partial x_2}\right)\right] = \frac{\partial \underline{v}}{\partial x_1} \cdot \left(\frac{\partial \underline{w}}{\partial x_2} \times \underline{w}\right) + \frac{\partial \underline{v}}{\partial x_2} \cdot \left(\underline{w} \times \frac{\partial \underline{w}}{\partial x_1}\right).$$

The main step to get the result is to use the divergence theorem appropriately and to see how that can be done first let

$$\underline{a} = \frac{\partial \underline{w}}{\partial x_2} \times \underline{w} = a_1\underline{e}_1 + a_2\underline{e}_2 + a_3\underline{e}_3 \quad \text{and} \quad \underline{b} = \underline{w} \times \frac{\partial \underline{w}}{\partial x_1} = b_1\underline{e}_1 + b_2\underline{e}_2 + b_3\underline{e}_3.$$

With v_1, v_2 and v_3 being the components of $\underline{v}$, the integrand in the right hand side of (3.4.19) can be written as

$$(a_1\underline{e}_1 + b_1\underline{e}_2)\cdot\nabla v_1 + (a_2\underline{e}_1 + b_2\underline{e}_2)\cdot\nabla v_2 + (a_3\underline{e}_1 + b_3\underline{e}_2)\cdot\nabla v_3.$$

As $\underline{v} = \underline{0}$ on the boundary $\partial\Omega$ the divergence theorem gives

$$\iint_\Omega (a_k\underline{e}_1 + b_k\underline{e}_2)\cdot\nabla v_k\, dx_1dx_2 = -\iint_\Omega v_k\left(\frac{\partial a_k}{\partial x_1} + \frac{\partial b_k}{\partial x_2}\right) dx_1dx_2, \quad k = 1,2,3.$$

Summing these results for $k = 1,2,3$ gives

$$\begin{aligned} I &= -\iint_\Omega \underline{v}\cdot\left[\frac{\partial}{\partial x_1}\left(\frac{\partial \underline{w}}{\partial x_2}\times\underline{w}\right) + \frac{\partial}{\partial x_2}\left(\underline{w}\times\frac{\partial \underline{w}}{\partial x_1}\right)\right] dx_1dx_2 \\ &= -\iint_\Omega \underline{v}\cdot\left[\left(\frac{\partial^2\underline{w}}{\partial x_1\partial x_2}\times\underline{w} + \frac{\partial \underline{w}}{\partial x_2}\times\frac{\partial \underline{w}}{\partial x_1}\right) + \left(\underline{w}\times\frac{\partial^2\underline{w}}{\partial x_1\partial x_2} + \frac{\partial \underline{w}}{\partial x_2}\times\frac{\partial \underline{w}}{\partial x_1}\right)\right] dx_1dx_2, \\ &= -2\iint_\Omega \underline{v}\cdot\left(\frac{\partial \underline{w}}{\partial x_2}\times\frac{\partial \underline{w}}{\partial x_1}\right) dx_1dx_2 = 2\iint_\Omega \underline{v}\cdot\left(\frac{\partial \underline{w}}{\partial x_1}\times\frac{\partial \underline{w}}{\partial x_2}\right) dx_1dx_2. \end{aligned}$$

In the above the property of the cross product that interchanging the terms changes the sign is used two times.

3.5 Some hyperelastic constitutive relations for the incompressible case

A hyperelastic material is a type of constitutive model for ideally elastic material which uses a strain energy function W for the stress-strain relationship, this is how the term is defined in Wikipedia. Ronald Rivlin and Melvin Mooney developed the first hyperelastic models, the neo-Hookean and Mooney Rivlin solids. Many other hyperelastic models have since been developed. A widely used way of representing W is the Ogden form which we use throughout this thesis. The expression of these models are given shortly and more details can be found in [4],[24], [21].

3.5.1 Stress-strain relations

The strain energy function W depends only on the deformation of the given model. Throughout this thesis we only consider isotropic and incompressible materials and in these cases W can be expressed in terms of the strain invariants I_1, I_2 given in (3.3.29) and (3.3.30). It can also be expressed in terms of the principal stretches λ_1 and λ_2. In

some situations that we need it is also convenient to consider W in other ways as for example a function of the full 3D deformation gradient $\mathbf{F}_{3D}$, or just of the membrane deformation gradient $\mathbf{F}$ and possibly of all 3 stretch ratios λ_1, λ_2 and λ_3. The meaning of partial derivatives depends on which version we are using and if we keep the same letter W for each of the functions then we can write

$$W = W(\mathbf{F}) = W(\mathbf{F}_{3D}) = W(I_1, I_2) = W(\lambda_1, \lambda_2, \lambda_3) = W(\lambda_1, \lambda_2). \tag{3.5.1}$$

We consider next the stress-stretch relations for our incompressible case for these different ways of representing W.

- **The strain energy function $W = W(\mathbf{F}_{3D})$**
  With W considered as a function of all 9 components of $\mathbf{F}_{3D}$ we have

  $$\mathbf{\Pi}_{3D}^T = \frac{\partial W}{\partial \mathbf{F}_{3D}}. \tag{3.5.2}$$

  This notation means that in components

  $$(\mathbf{\Pi}_{3D}^T)_{ij} = \frac{\partial W}{\partial (\mathbf{F}_{3D})_{ij}}.$$

- **The strain energy function $W = W(\mathbf{F})$**
  With W considered as a function of the 6 components of the membrane deformation gradient $\mathbf{F}$ for the incomprerssible membrane deformations being considered it can be shown that

  $$\mathbf{\Pi}^T = \frac{\partial W}{\partial \mathbf{F}}. \tag{3.5.3}$$

  Unlike in other situations where the stress is not completely determined by the deformation when the deformation is incompressible it is in the membrane case as a consequence of the membrane assumption. We do not give details here to explain this further here but they are similar to the details given below when we consider how to obtain the version involving $W = W(\lambda_1, \lambda_2)$ from the version involving $W = W(\lambda_1, \lambda_2, \lambda_3)$. Accepting that we have (3.5.3) it follows that the stress term (3.4.16) in the weak form of the membrane case can be expressed as follows

  $$\iint_\Omega \mathbf{\Pi}^T : \nabla \underline{v}\, dx_1 dx_2 = \iint_\Omega \frac{\partial W}{\partial \mathbf{F}} : \nabla \underline{v}\, dx_1 dx_2. \tag{3.5.4}$$

  We note here that the above expression is used in our implementations for the pressure model.

- **The strain energy function** $W = W(I_1, I_2)$
  In several text books the incompressible and isotropic case is described with $W = W(I_1, I_2)$. To give the stress-stretch relation in a compact form we use the following notation for the first partial derivatives.

$$W_1 = \frac{\partial W}{\partial I_1} \quad \text{and} \quad W_2 = \frac{\partial W}{\partial I_2}. \tag{3.5.5}$$

  The constitutive equation can be expressed in either of the following forms

$$\boldsymbol{\sigma}_{3D} = -pI + 2(W_1 + I_1 W_2)\mathbf{B}_{3D} - 2W_2\mathbf{B}_{3D}^2 \tag{3.5.6}$$

$$\boldsymbol{\sigma}_{3D} = -pI + 2W_1\mathbf{B}_{3D} - 2W_2\mathbf{B}_{3D}^{-1} \tag{3.5.7}$$

  with, as before, $\boldsymbol{\sigma}_{3D}$ being the general 3D version of the Cauchy Stress tensor and $\mathbf{B}_{3D}$ being the **Left Cauchy Green deformation tensor** as it was described in (3.3.5). In these relations p is a hydrostatic pressure term as a consequence of the incompressibility assumption and for a general incompressible deformation p has to be determined as part of computation. The proof of the above form can be found in [29, p.141]. We do not repeat the details of the proof but it is worth commenting on why an incompressible deformation does not completely determine the stress for a general 3D hyperelastic body. The origins of this is a power balance requirement for a body to be elastic with in addition the property that we have a hyperelastic assumption with a strain energy function W. The power balance requirement is that

$$\rho\frac{\mathrm{d}W}{\mathrm{d}t} = \boldsymbol{\sigma}_{3D} : \mathbf{D}, \tag{3.5.8}$$

  where ρ is the deformed density and $\mathbf{D}$ is the rate of deformation tensor, i.e.

$$\mathbf{D} = (D_{ij}), \quad D_{ij} = \frac{1}{2}\left(\frac{\partial \dot{w}_i}{\partial w_j} + \frac{\partial \dot{w}_j}{\partial w_i}\right).$$

  Here $\underline{w} = \underline{x} + \underline{u}$ is the deformed position and $\underline{\dot{w}} = (\dot{w}_i)$ is the velocity. The term $\boldsymbol{\sigma}_{3D} : \mathbf{D}$ is known as the stress power. For non-elastic materials we do not have a W and there are other terms in (3.5.8) concerned with dissipation. When a material can only deform in an incompressible way we have $\det \mathbf{F}_{3D} = 1$ and this can be equivalently written in the form

$$\nabla \cdot \underline{\dot{w}} = \operatorname{tr} \mathbf{D} = \frac{\partial \dot{w}_1}{\partial w_1} + \frac{\partial \dot{w}_2}{\partial w_2} + \frac{\partial \dot{w}_3}{\partial w_3} = 0.$$

  When the only possible deformations are incompressible deformations the condition (3.5.8) is unchanged if we add or subtract any multiple of the identity $\mathbf{I}$ from

an expression for $\boldsymbol{\sigma}_{3D}$, That is, $\boldsymbol{\sigma}_{3D}$ is only determined up to the addition of a term such as $-p\mathbf{I}$. The derivation in [29] to get the expressions given in (3.5.6) and (3.5.7) is to first derive the relations for general I_1, I_2 and I_3 and then to set things up to get the form of the equations in the limit $I_3 \to 1$. In the case of a membrane model we have $\boldsymbol{\sigma} \approx \boldsymbol{\sigma}_{3D}$ evaluated on the mid-surface with $\boldsymbol{\sigma}$ being such that $\boldsymbol{\sigma}\underline{n} = \underline{0}$. As this must be satisfied the term p in (3.5.6) and (3.5.7) is determined in terms of other quantities, i.e. by using (3.4.4) and (3.4.10) we have

$$0 = \left(\mathbf{R}^T\boldsymbol{\sigma}\mathbf{R}\right) = -p + 2W_1\lambda^2 - 2W_2\lambda^{-2}, \tag{3.5.9}$$

which implies that

$$p = 2(W_1\lambda^2 - W_2\lambda^{-2}).$$

- **The strain energy function** $W = W(\lambda_1, \lambda_2, \lambda_3)$
  In the incompresible case

$$\sigma_1 = -p + \lambda_1\frac{\partial W}{\partial \lambda_1}, \quad \sigma_2 = -p + \lambda_2\frac{\partial W}{\partial \lambda_2}, \quad \sigma_3 = -p + \lambda_3\frac{\partial W}{\partial \lambda_3} \tag{3.5.10}$$

  where, as in the previous case, p is a hydrostatic pressure.

- **The strain energy function** $W = W(\lambda_1, \lambda_2)$
  This is the version that we usually use when describing a strain energy function and leads to the two non-zero principal stresses σ_1 and σ_2 being given by

$$\sigma_1 = \lambda_1\frac{\partial W}{\partial \lambda_1} \quad \text{and} \quad \sigma_2 = \lambda_2\frac{\partial W}{\partial \lambda_2} \tag{3.5.11}$$

  for our incompressible and isotropic membrane deformation. To understand how these relations follow from (3.5.10) can be done as follows.

  It is convenient here to distinguish between the cases by using the letter W for this case and $\widetilde{W}$ for the previous case with

$$W(\lambda_1, \lambda_2) = \widetilde{W}(\lambda_1, \lambda_2, \lambda_3) = \widetilde{W}(\lambda_1, \lambda_2, 1/(\lambda_1\lambda_2)). \tag{3.5.12}$$

  As $\sigma_3 = 0$ we have

$$p = \lambda_3\frac{\partial \widetilde{W}}{\partial \lambda_3}. \tag{3.5.13}$$

  For the other two principal stresses we have the following. By taking the derivative

of $W = W(\lambda_1, \lambda_2)$ with respect to λ_1 and λ_2 respectively we get

$$\frac{\partial W}{\partial \lambda_1} = \frac{\partial \widetilde{W}}{\partial \lambda_1} + \frac{\partial \widetilde{W}}{\partial \lambda_3}\frac{\partial \lambda_3}{\partial \lambda_1} = \frac{\partial \widetilde{W}}{\partial \lambda_1} + \frac{\partial \widetilde{W}}{\partial \lambda_3}(-\lambda_1^{-2}\lambda_2^{-1}), \tag{3.5.14}$$

$$\frac{\partial W}{\partial \lambda_2} = \frac{\partial \widetilde{W}}{\partial \lambda_2} + \frac{\partial \widetilde{W}}{\partial \lambda_3}\frac{\partial \lambda_3}{\partial \lambda_2} = \frac{\partial \widetilde{W}}{\partial \lambda_2} + \frac{\partial \widetilde{W}}{\partial \lambda_3}(-\lambda_1^{-1}\lambda_2^{-2}). \tag{3.5.15}$$

Then by multiplying each of these equations (3.5.14) and (3.5.15) by λ_1 and λ_2 respectively we get the following

$$\lambda_1\frac{\partial W}{\partial \lambda_1} = \lambda_1\frac{\partial \widetilde{W}}{\partial \lambda_1} - \lambda_3\frac{\partial \widetilde{W}}{\partial \lambda_3}, \tag{3.5.16}$$

$$\lambda_2\frac{\partial W}{\partial \lambda_2} = \lambda_2\frac{\partial \widetilde{W}}{\partial \lambda_2} - \lambda_3\frac{\partial \widetilde{W}}{\partial \lambda_3}. \tag{3.5.17}$$

Therefore, by combining the equations (3.5.10) with the quantities (3.5.16), (3.5.17) and (3.5.13) we have the following expressions for the principal stresses σ_1 and σ_2 in terms of λ_1 and λ_2 respectively

$$\sigma_1 = -p + \lambda_1\frac{\partial \widetilde{W}}{\partial \lambda_1} = -\lambda_3\frac{\partial \widetilde{W}}{\partial \lambda_3} + \lambda_1\frac{\partial \widetilde{W}}{\partial \lambda_1} = \lambda_1\frac{\partial W}{\partial \lambda_1}, \tag{3.5.18}$$

$$\sigma_2 = -p + \lambda_2\frac{\partial \widetilde{W}}{\partial \lambda_2} = -\lambda_3\frac{\partial \widetilde{W}}{\partial \lambda_3} + \lambda_2\frac{\partial \widetilde{W}}{\partial \lambda_2} = \lambda_2\frac{\partial W}{\partial \lambda_2}. \tag{3.5.19}$$

3.5.2 Examples of strain energy functions

We now give some examples of hyperelastic models by using strain energy functions in terms of I_1, I_2 and/or λ_1, λ_2.

In terms of I_1 and I_2 we have the neo-Hookean and Mooney-Rivlin models, which respectively have the following representations. The neo-Hookean model is given by

$$W = C(I_1 - 3), \quad \text{where } C > 0 \text{ is a constant} \tag{3.5.20}$$

The Mooney-Rivlin model is given by

$$W = C((I_1 - 3) + a(I_2 - 3)), \quad \text{where } a, C \text{ are constants.} \tag{3.5.21}$$

In terms of λ_1 and λ_2 the Ogden form representation which involves $W = W(\lambda_1, \lambda_2)$

is an expression of the form

$$W = \sum_{1}^{N} \frac{\mu_p}{\nu_p}(\lambda_1^{\nu_p} + \lambda_2^{\nu_p} + \lambda_1^{-\nu_p}\lambda_2^{-\nu_p} - 3), \quad \frac{\mu_p}{\nu_p} > 0 \tag{3.5.22}$$

where μ_p and ν_p are constants. When W is written in the Ogden form it is easy to compute the principal stresses σ_1 and σ_2 by using (3.5.11). An example of this form is the Jones-Treloar hyperelastic model which has the following representation

$$\begin{aligned} W &= \frac{0.69}{1.3}(\lambda_1^{1.3} + \lambda_2^{1.3} + \lambda_1^{-1.3}\lambda_2^{-1.3} - 3) \\ &+ \frac{0.01}{4}(\lambda_1^4 + \lambda_2^4 + \lambda_1^{-4}\lambda_2^{-4} - 3) + \\ &+ \frac{-0.0122}{-2}(\lambda_1^{-2} + \lambda_2^{-2} + \lambda_1^2\lambda_2^2 - 3). \end{aligned} \tag{3.5.23}$$

In addition the neo-Hookean and Mooney Rivlin models given above can be also expressed by using the Ogden form. For the neo-Hookean model given above we have

$$W = C(\lambda_1^2 + \lambda_2^2 + \lambda_3^2 - 3) \tag{3.5.24}$$

and for the Mooney Rivlin model we have

$$W = C((\lambda_1^2 + \lambda_2^2 + \lambda_3^2 - 3) + a(\lambda_1^2\lambda_2^2 + \lambda_1^2\lambda_3^2 + \lambda_2^2\lambda_3^2 - 3)). \tag{3.5.25}$$

The above presented hyperelastic models are used in this thesis when computational results are presented.

3.6 The numerical scheme and the implementation of the pressure model

In this section we describe aspects of the finite element method to attempt to get an approximation $\underline{u}_h$ in a finite element space V_h of the displacement field $\underline{u}$. The finite element spaces used in this thesis involve piecewise polynomials of degree 1 or 2 defined on a triangular mesh of Ω and give functions which are in $C(\bar{\Omega})$, i.e. they are continuous in $\bar{\Omega}$, and are such that they are in $H^1(\Omega)$. The finite element test functions used are similarly in this space and they also vanish on the boundary $\partial\Omega$. The detail given is of the nonlinear equations involved and the computation that is required on each element to get what we call the element residual and the element Jacobian matrix which, when assembled, give the vector valued function in the nonlinear equations and the Jacobian matrix of that function. Both of these are needed in a Newton iteration. As we show the

Jacobian matrix is symmetric. As we also explain, care is needed at some stages to do the computations in such a way that we avoid division by 0 in some intermediate steps.

3.6.1 A prestretch and getting a solution for the first non-zero pressure

In this section we briefly comment on the practical aspect of getting the solution at the first non-zero pressure and leave to subsection 3.6.6 further theoretical comments to justify things a bit further. The theoretical comments are done in a later section as they need some of the expressions which are given in the next few subsections where Newton's method is described.

The details in these subsections are about the numerical scheme to approximately solve the following nonlinear problem. Find the displacement field $\underline{u}$ from the appropriate space which satisfies

$$\mathcal{A}(\underline{u};\underline{v},P)=0, \quad \text{where } \mathcal{A}(\underline{u};\underline{v},P)=a_1(\underline{u};\underline{v})-Pa_2(\underline{u};\underline{v}), \tag{3.6.1}$$

for all $\underline{v}$ from the appropriate space. The term $\mathcal{A}(\underline{u};\underline{v},P)$ is the same as what we usually write as $A(\underline{u};\underline{v})$ and is used here to indicate explicitly that the solution $\underline{u}$ depends on the pressure P. The flat undeformed starting state corresponding to $\underline{u}=\underline{0}$ is the solution throughout the domain Ω when the pressure $P=0$ and we have $\underline{u}=\underline{0}$ on $\partial\Omega$ as the boundary condition. If we give the flat sheet a uniform prestretch, e.g.

$$u_1(x_1,x_2)=\beta x_1, \quad u_2(x_1,x_2)=\beta x_2, \quad u_3(x_1,x_2)=0, \tag{3.6.2}$$

where $\beta>0$ is a constant, then we also have a solution (3.6.1) when $P=0$ but instead in the case of appropriate non-zero boundary conditions on $\partial\Omega$ corresponding to the prestretch values. As we show in subsequent sections, the Jacobian matrix of the nonlinear equations that we have to solve is symmetric and there is a term which comes from the $a_1(.;.)$ part and there is a term which comes from the $a_2(.;.)$ part. When we start from a prestretched state, i.e. $\beta>0$, the term which comes from the $a_1(.;.)$ term gives a symmetric positive definite matrix whilst the term which comes from the $a_2(.;.)$ part is just symmetric. For small values of P the Jacobian matrix is hence symmetric and positive definite when we evaluate it at a displacement field corresponding to the prestretched state. The positive definite property of the Jacobian matrix term which arises from the $a_1(.;.)$ term is as a consequence of the linearised problem to get the solution close to the prestretched state. This is a linear problem which has a coercive property and we give some details about this in subsection 3.6.6. If we want to solve the problem corresponding

to $\underline{u} = \underline{0}$ on $\partial\Omega$, i.e. we do not have a prestretch, then the contribution to the Jacobian matrix from the $a_1(.;.)$ term is a singular matrix when we evaluate at a displacement field corresponding to $\underline{u} = 0$ everywhere. For a little more detail here, the rows and columns corresponding to every component of $(\underline{u})_3$ are identically zero and we give a few more comments about this in subsection 3.6.6. Hence when we have a prestretch we can take this as the starting point in a Newton iteration for a small pressure $P_1 > 0$ to attempt to get a displacement field corresponding to the flat prestretched state but we cannot do this when we want the solution corresponding to $\beta = 0$.

There are no results in this thesis for the inflation problem corresponding to $\underline{u} = 0$ on $\partial\Omega$ as the boundary condition, this corresponds to having no prestretch, but there is a strategy involving not too many additional steps to deal with this situation. To describe this we briefly use the notation $\underline{u}(\underline{x}, P_1, \beta)$ to denote the solution with pressure P_1 when the boundary conditions correspond to a prestretch β. To attempt to get $\underline{u}(\underline{x}, P_1, 0)$ we first obtain $\underline{u}(\underline{x}, P_1, \beta)$ for some small $\beta \neq 0$. We then create a displacement field $\underline{\tilde{u}}(\underline{x}, P_1, 0)$ from $\underline{u}(\underline{x}, P_1, 0)$ which is such that $\underline{\tilde{u}}(\underline{x}, P_1, 0) = \underline{0}$ when $\underline{x} \in \partial\Omega$. This vector will not satisfy the weak problem but it can be used as the starting vector in the Newton iteration to attempt to get $\underline{u}(\underline{x}, P_1, 0)$. How we get a possible vector $\underline{\tilde{u}}(\underline{x}, P_1, 0)$ depends on the shape of Ω. In the case of $\Omega = \{(x_1, x_2) : -1 \leq x_1, x_2 \leq 1\}$ we can for example define $\underline{\tilde{u}}$ via the equation

$$\underline{x} + \underline{\tilde{u}}(\underline{x}, P_1, 0) = \left(\frac{1}{1+\beta}\right)(\underline{x} + \underline{u}(\underline{x}, P_1, \beta)). \tag{3.6.3}$$

If the iteration does not converge then we can use $\underline{u}(\underline{x}, P_1, \beta)$ to attempt to get the solution $\underline{u}(\underline{x}, P_1, \tilde{\beta})$ for some $0 < \tilde{\beta} < \beta$ in a similar manner to the above. Provided the solution varies continuously with the parameter $\tilde{\beta}$ and every Jacobian matrix on the path is non-singular we can use a continuation strategy to attempt to get the desired solution $\underline{u}(\underline{x}, P_1, 0)$.

3.6.2 The equations for different pressures $0 = P_0 < P_1 < \ldots$

In this section we assume that we start with a prestretched state corresponding to $u_3 = 0$ and $\mathbf{F} \neq \mathbf{I}$ being constant throughout the domain and as indicated in the previous section this is the solution at the pressure $P_0 = 0$.

Suppose that we have a mesh of Ω which gives us the finite element space involving piecewise linear elements and suppose that the spaces can be described as

$$\underline{u}_h \in \text{span}\{\underline{v}_1, \ldots, \underline{v}_m\} \quad \text{and} \quad V_h = \text{span}\{\underline{v}_1, \ldots, \underline{v}_{\tilde{m}}\}, \quad \tilde{m} < m \tag{3.6.4}$$

with the functions with index values $\tilde{m}+1,\ldots,m$ corresponding to points on $\partial\Omega$ where $\underline{u}_h$ is known. At any given pressure P_k the finite element problem involves determining $\underline{u}_h$ such that

$$A_i(\underline{u}_h;P_k)=a_1(\underline{u}_h;\underline{v}_i)-P_k a_2(\underline{u}_h;\underline{v}_i)=\underline{0},\quad i=1,\ldots,\tilde{m}. \tag{3.6.5}$$

Let $\underline{A}=(A_i)$ and let $\underline{c}^{(j)}$ denote the unknown parameters of a candidate function $\underline{u}_h^{(j)}$ in a Newton iteration to attempt to solve (3.6.5) with the Newton iteration being of the form

$$\underline{c}^{(j+1)}=\underline{c}^{(j)}-J_A(\underline{c}^{(j)};P_k)^{-1}\underline{A}(\underline{u}_h^{(j)};P_k),\quad j=0,1,2,\ldots \tag{3.6.6}$$

where $J_A(\underline{c}^{(j)};P_k)$ is the Jacobian matrix associated with $\underline{A}(\underline{u}_h^{(j)};P_k)$ in terms of how it depends on $\underline{c}^{(j)}$. The element-by-element details of constructing contributions to $\underline{A}$ and J_A are discussed later in section 3.6.3 and this might be considered as the finer details. It is also important to have an overall set-up so that it is likely that the Newton iteration in (3.6.6) will converge when the system of equations has a solution. This can usually be done by considering the problem at a sequence of pressures

$$0=P_0<P_1<\cdots<P_k<\cdots$$

If P_1 is close to P_0 then the solution at P_0 is likely to be close to the solution at P_1 and this generates a good starting vector for the Newton iteration. This approach is repeated in an attempt to get a solution at pressure P_k given that we have obtained a solution at pressure P_{k-1}. If the Newton iteration does not converge at pressure P_k then we replace P_k by a pressure closer to P_{k-1}. We summarize this next as an algorithm when we attempt increase the pressure in steps of magnitude P_{step}.

Algorithm

1. Let $k=1$, $P_0=0$ and get the prestretch state to be used.
2. Set $P_k=P_{k-1}+P_{step}$.
3. Attempt to find the approximate solution $\underline{u}_h$ with the pressure P_k starting the Newton iteration with the solution at P_{k-1}.
4. If the iteration converges then replace k by $k+1$.
5. If the iteration does not converge then replace P_{step} by $P_{step}/2$.
6. Goto step 2 or stop if all the required solutions have been obtained.

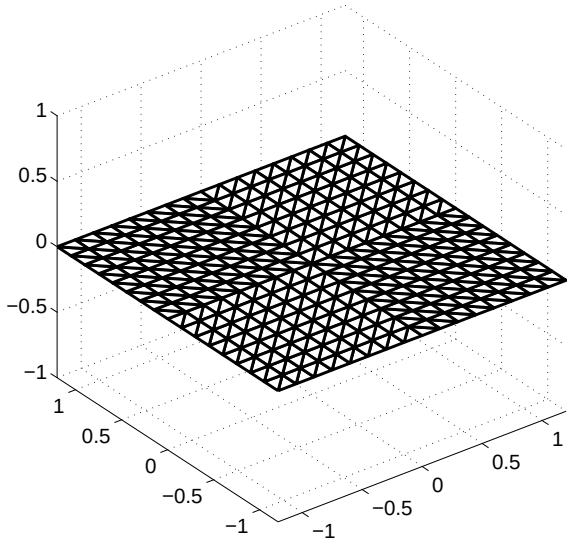
This procedure works when the geometry and the constitutive model are such that there is a solution and in the following we show graphically the deformed membranes which are obtained in a few cases. In the case of a Mooney Rivlin model corresponding to $W = C((I_1 - 3) + a(I_2 - 3))$ with $C = 0.5$ $a = 0.1$ we show in figure 3.1 the flat sheet at the start, an intermediate state and the deformed sheet at the final pressure considered which gives a height at the centre which is just above 2. At the start of the deformation there is a prestretch of 0.2 and the mesh is a uniform mesh of 512 elements which involves 289 nodes. A prestretch of magnitude 0.2 is used in all the examples shown here. In the case of a neo-Hookean model corresponding to $W = C(I_1 - 3)$ with $C = 1$ we have a mesh which is approximately a 2:1 ellipse and has 311 elements and 180 nodes. The starting state and the deformed sheet when the maximum height is just above 1 are shown in Figure 3.2(a). In the case of the Jones Treloar model given in (3.5.23) we have a mesh which is approximately the unit circle and which has 148 elements and 88 nodes. The starting state and the state when the maximum height is just above 1.2 are shown in Figure 3.3. In the case of the Jones Treloar model again we have a mesh of a square with a circular hole which has 413 elements and 251 nodes. The starting state and the state when the maximum height is just above 0.6 are shown in Figure 3.4. It is interesting to compare the effect of doing essentially the same problem with the different hyperelastic models and we do this in the case of the mesh used in Figure 3.1 and with the pressure in each example adjusted so that we

$$(\underline{u}_h)_3(0,0) - H = 0, \quad \text{with } H = 1.2. \tag{3.6.7}$$

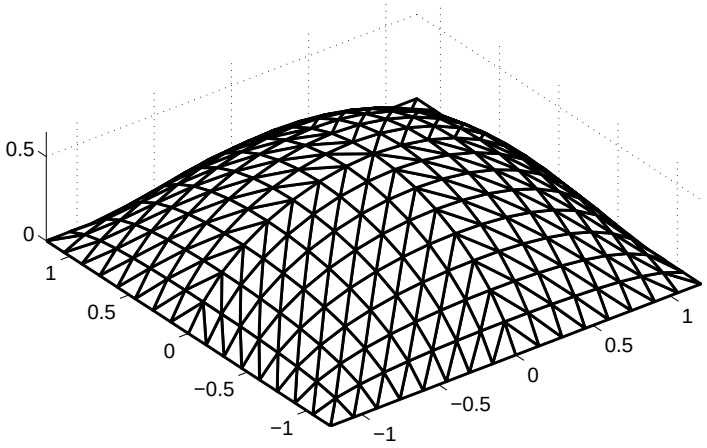
This is done by adding (3.6.7) to the set of equations to be solved and with adding the pressure as an additional unknown. The 3 different deformed states are shown in Figure 3.5, where we consider a square sheet. It is difficult when presented in this way to easily detect that the deformed states are different and if you consider the differences between the $\underline{u}_h$ values at the 289 nodes the greatest difference is only about 0.047.

3.6.3 Computational details on the element level

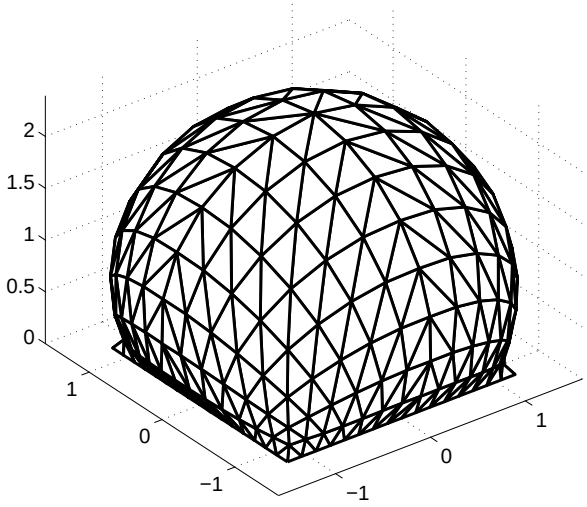
In this section we describe some of the computations that we need to do on an element to compute the residual $\underline{A}$ and the Jacobian matrix J_A which appeared in (3.6.6). To simplify the notation here we just write $\underline{u}_h$ instead of $\underline{u}_h^{(j)}$. On an element we let $\underline{\phi}_i$ denote a basis function and we now let m denote the number of functions needed on the element.



(a) Rectangle: Starting Shape

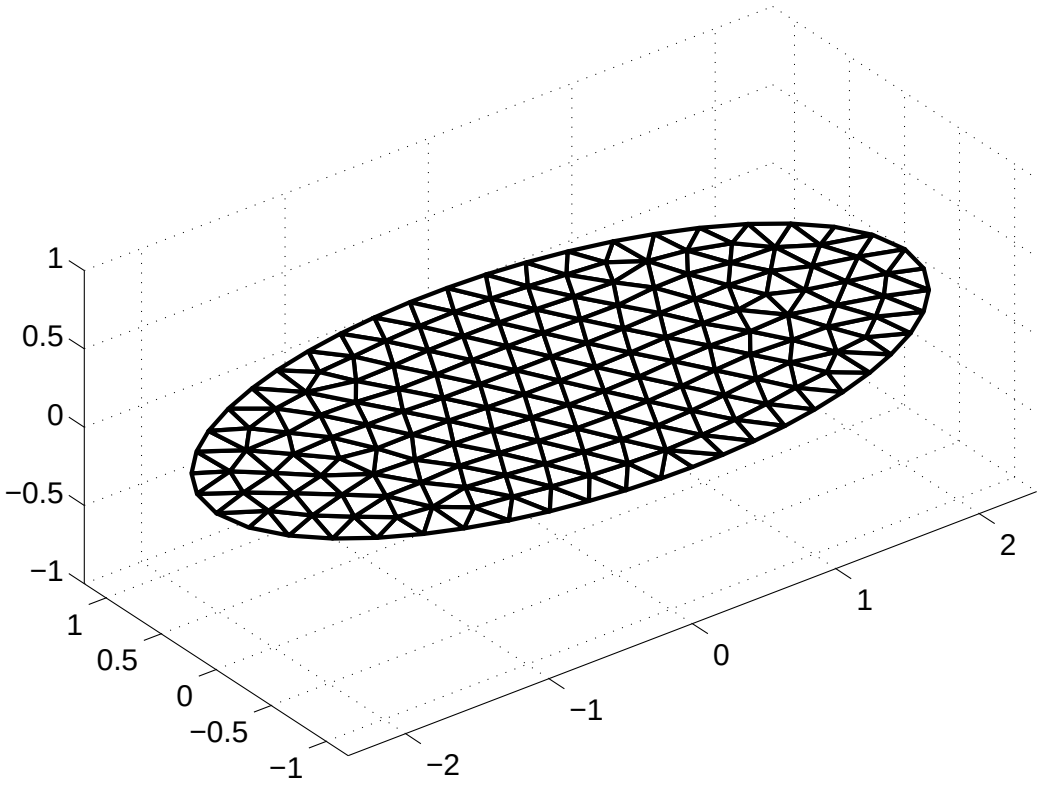


(b) Rectangle: Intermediate Shape

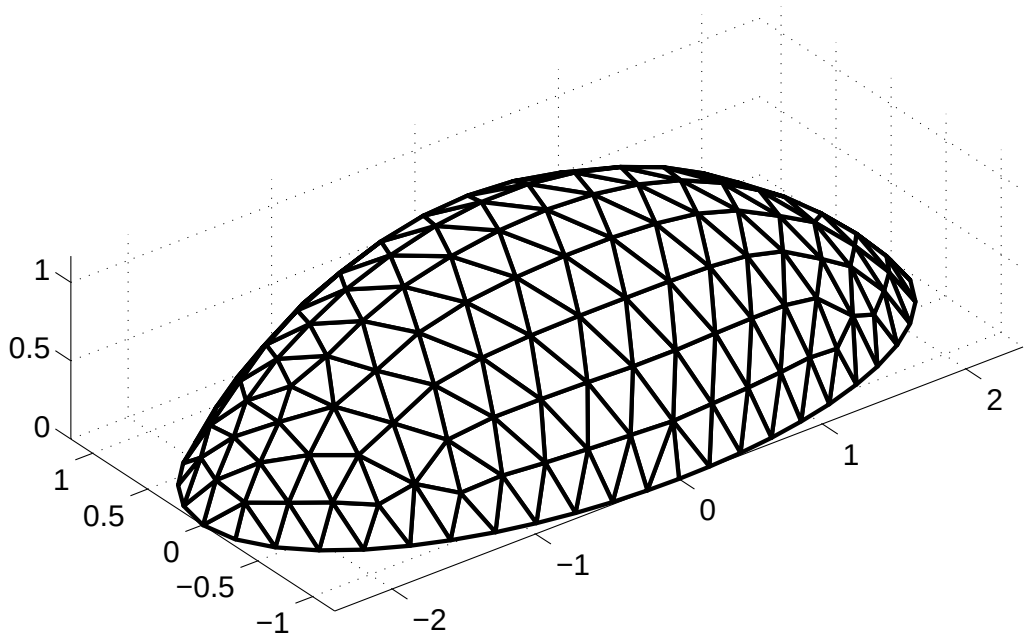


(c) Rectangle: Final shape

Fig. 3.1: Deforming a membrane in the shape of a square

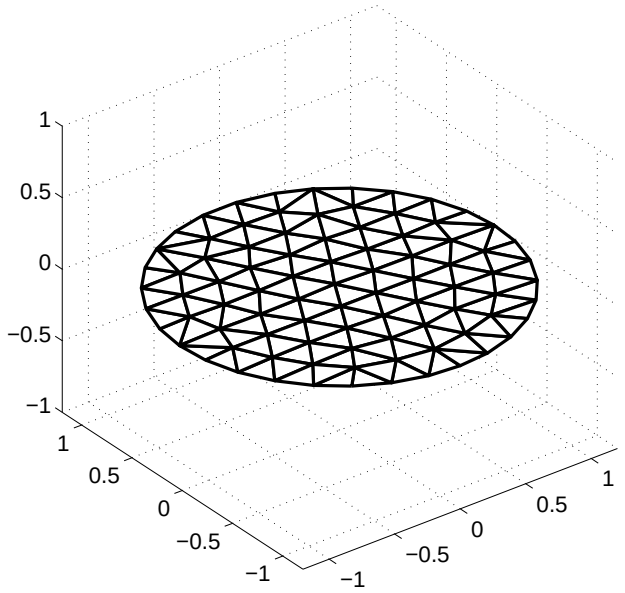


(a) Ellipse mesh at the start

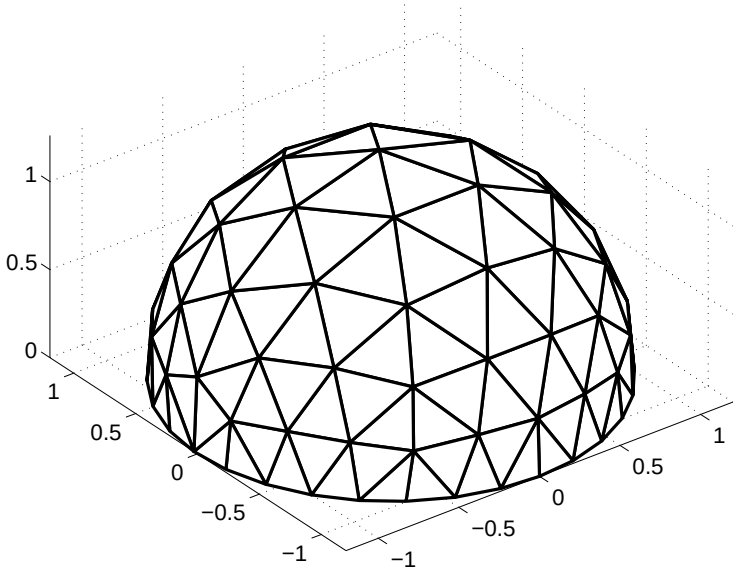


(b) Deformed ellipse

Fig. 3.2: The triangular mesh approximately represents a 2:1 ellipse. Fig. 3.2(a) show the starting state and Fig. 3.2(b) show the deformed state when the pressure is such that the height at the centre is just over 1

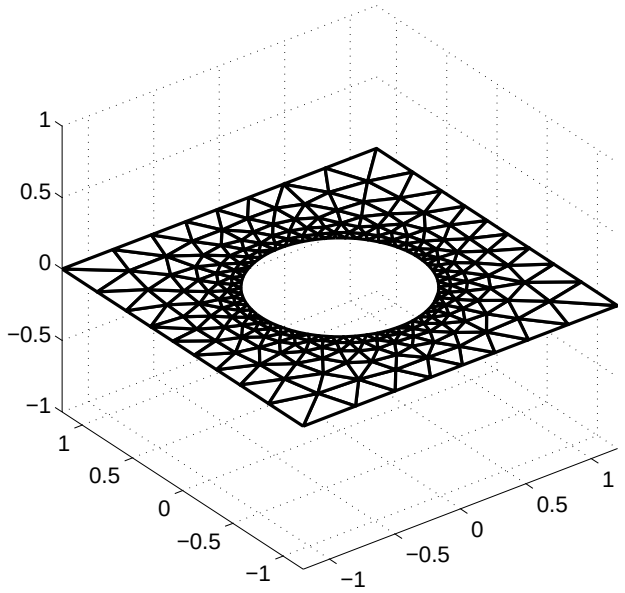


(a) Circle mesh at the begin

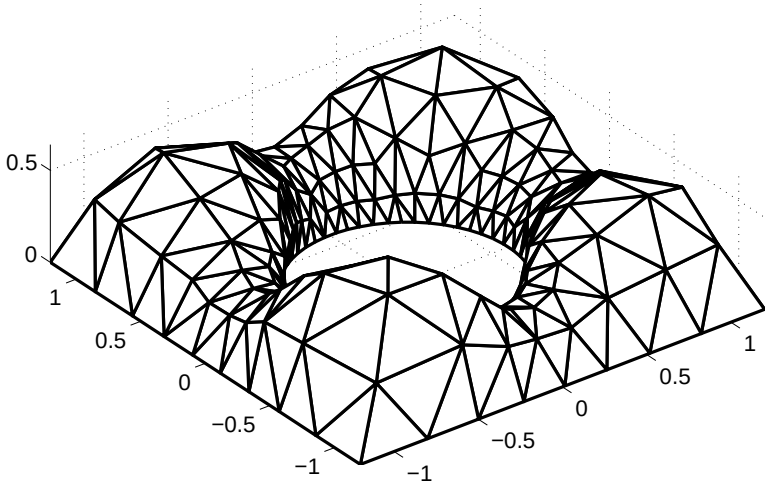


(b) Circle mesh at the end

Fig. 3.3: Deforming a membrane in the shape of a circle

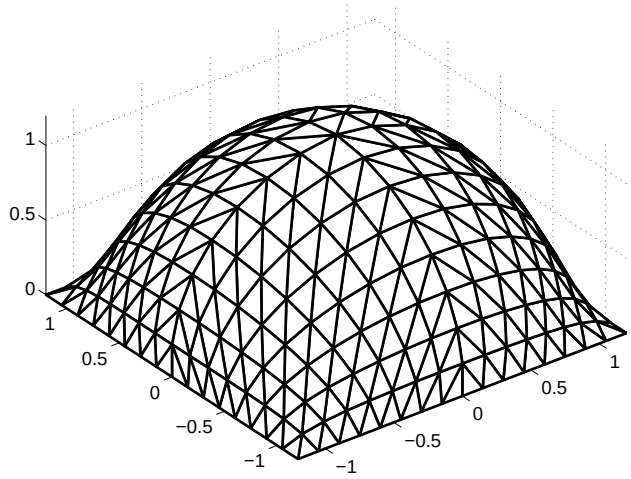


(a) Rectangle with circular hole at the start

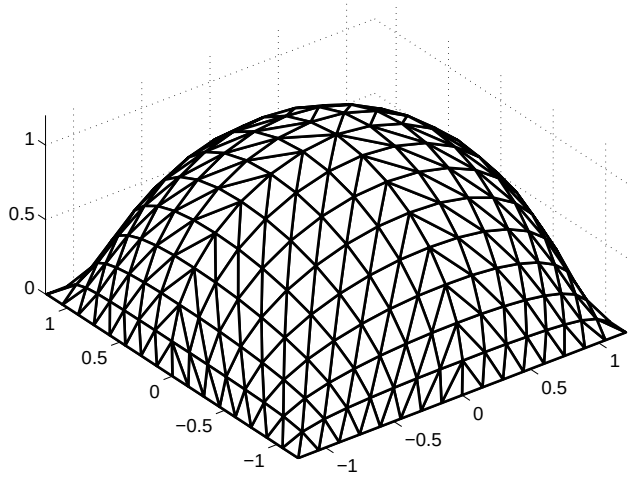


(b) Rectangle with circular hole at the end

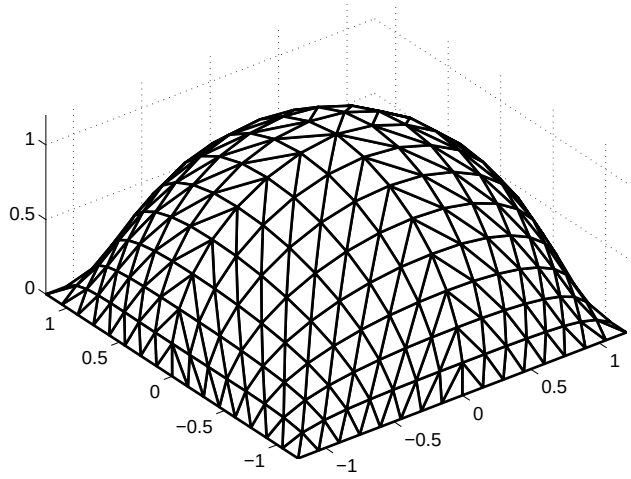
Fig. 3.4: Deforming a membrane in the shape of a rectangle with a hole



(a) Neo-Hookean model, the height at the centre is 1.2



(b) Mooney Rivlin model, the height at the centre is 1.2



(c) Jones Treloar model, the height at the centre is 1.2

Fig. 3.5: In each case the same mesh is used and the pressure P is determined so that the height at the centre is 1.2.

The finite element function $\underline{u}_h$ has the form

$$\underline{u}_h = \sum_{i=1}^{m} u_i \underline{\phi}_i. \tag{3.6.8}$$

We consider triangles throughout and as there are 3 displacement components at each point we have $m = 9$ with linear elements and $m = 18$ with quadratic elements. For the description here we just consider the case $m = 9$. To describe a triangle we use similar notation as it was used in chapter 2 and we let $\underline{x}_1, \underline{x}_2, \underline{x}_3$ denote the three nodes of a triangle Ω_r and we let N_1, N_2, N_3 denote the scalar basis functions. Associated with these 3 functions we have the 9 vector valued functions $\underline{\phi}_1, \underline{\phi}_2, \underline{\phi}_3, \ldots, \underline{\phi}_9$ as indicated by

$$(\underline{\phi}_1, \underline{\phi}_2, \underline{\phi}_3, \ldots, \underline{\phi}_9) = \begin{pmatrix} N_1 & 0 & 0 & N_2 & 0 & 0 & N_3 & 0 & 0 \\ 0 & N_1 & 0 & 0 & N_2 & 0 & 0 & N_3 & 0 \\ 0 & 0 & N_1 & 0 & 0 & N_2 & 0 & 0 & N_3 \end{pmatrix}. \tag{3.6.9}$$

These represent the test vectors on the element. Corresponding to this notation we present the nodal displacement values in a similar way. We let $\underline{u}_h(\underline{x}_1), \underline{u}_h(\underline{x}_2), \underline{u}_h(\underline{x}_3)$ be the displacements at the nodes and then we define the corresponding vector

$$\underline{u} = (u_i) = ((\underline{u}_h(\underline{x}_1))_1, (\underline{u}_h(\underline{x}_1))_2, (\underline{u}_h(\underline{x}_1))_3, (\underline{u}_h(\underline{x}_2))_1, \ldots, (\underline{u}_h(\underline{x}_3))_3)^T \quad \in \mathbb{R}^9 \tag{3.6.10}$$

which has all the nodal displacement values. Using this notation the function $\underline{u}_h$ is given by

$$\underline{u}_h(x_1, x_2) = \sum_{i=1}^{9} u_i \underline{\phi}_i(x_1, x_2). \tag{3.6.11}$$

If we let $a_1(\cdot;\cdot)_{\Omega_r}$ and $a_2(\cdot;\cdot)_{\Omega_r}$ mean the usual expressions for $a_1(\cdot;\cdot)$ and $a_2(\cdot;\cdot)$ with instead the integrals taken over the element Ω_r, i.e.

$$a_1(\underline{u}_h; \underline{\phi}_i)_{\Omega_r} = h_0 \iint_{\Omega_r} \frac{\partial W}{\partial \mathbf{F}} : \nabla \underline{\phi}_i \, dx_1 dx_2, \tag{3.6.12}$$

$$\begin{aligned} a_2(\underline{u}_h; \underline{\phi}_i)_{\Omega_r} = {} & \frac{1}{3} \iint_{\Omega_r} \underline{\phi}_i \cdot \left(\frac{\partial \underline{w}}{\partial x_1} \times \frac{\partial \underline{w}}{\partial x_2} \right) dx_1 dx_2 + \frac{1}{3} \iint_{\Omega_r} \underline{w} \cdot \left(\frac{\partial \underline{\phi}_i}{\partial x_1} \times \frac{\partial \underline{w}}{\partial x_2} \right) dx_1 dx_2 \\ & + \frac{1}{3} \iint_{\Omega_r} \underline{w} \cdot \left(\frac{\partial \underline{w}}{\partial x_1} \times \frac{\partial \underline{\phi}_i}{\partial x_2} \right) dx_1 dx_2 \end{aligned} \tag{3.6.13}$$

then corresponding to the terms $A_i(\cdot;\cdot)$ which are in (3.6.5) the element version is defined by

$$A_i(\underline{u}_h; P_k)_r = a_1(\underline{u}_h; \underline{\phi}_i)_{\Omega_r} - P_k a_2(\underline{u}_h; \underline{\phi}_i)_{\Omega_r}, \quad 1 \le i \le 9. \tag{3.6.14}$$

On an element we thus have a 9×1 vector of nodal values and we have to compute a 9×1 element residual vector $\underline{A}(\underline{u}_h; P_k)_r = (A_i(\underline{u}_h; P_k)_r)$. To get the Jacobian matrix needed in (3.6.6) we also need to compute the 9×9 element Jacobian matrix which we denote by $J_A(\underline{u}_h; P_k)$. Some details about the element Jacobian matrix and why it is symmetric are given in the next section.

3.6.4 The element Jacobian matrix

In the following, as the function $\underline{u}_h$ is defined by its nodal values $\underline{u} \in \mathbb{R}^9$ we will write $a_1(\underline{u}; P_k)_r$ and $a_2(\underline{u}; P_k)_r$ for $a_1(\underline{u}_h; P_k)_r$ and $a_1(\underline{u}_h; P_k)_r$ respectively and similarly we will write $\underline{A}(\underline{u}; P_k)$ for $\underline{A}(\underline{u}_h; P_k)$. For the Jacobian element matrix $J_A(\underline{u}; P_k)_r$, we need to find the partial derivatives of $\underline{A}(\underline{u}; P_k)_r$ with respect to parameters $u_1, \ldots, u_9$.

Now, if you only wish to get the approximate solution and you are not so concerned about the efficiency then the element Jacobian matrix $J_{\underline{A}}(\underline{u}; P_k)_r$ can be approximated column-by-column by using finite differences. For example, in the case of the r^{th} column we have

$$J_A(\underline{u}; P_k)\underline{e}_r \approx \frac{\underline{A}(\underline{u} + h\underline{e}_j; P_k) - \underline{A}(\underline{u}; P_k)}{h} \quad \text{for small } h, \tag{3.6.15}$$

where $\underline{e}_r$ denotes the r^{th} column of the 9×9 identity matrix. Although this is a fairly simple thing to do it does involve evaluating $\underline{A}(\cdot; P_k)_r$ at 10 different displacements when we have linear triangles. The similar computation using the 6-noded quadratic triangles involves evaluating at 19 different displacements. It is better to get the expression for $J_A(\underline{u}; P_k)$ and, as we will see later, the same expression also appears when the Gâteaux derivatives are needed when dual problems are described.

The symmetry of the contribution from a_1

In the double dot product operation

$$\frac{\partial W}{\partial \mathbf{F}} : \nabla \underline{\phi}_i \tag{3.6.16}$$

both terms have shape 3×2 which is the shape of $\mathbf{F}$. It is convenient for the description, and also for the implementation, to have a version of these which are reshaped as 6×1 and to refer to the individual entries with a single index accordingly. In each case the 6 entries are the entries in the first column followed by the entries in the second column.

In the case of $\mathbf{F}$ we define $\underline{F}$ by

$$\mathbf{F} = \begin{pmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \\ F_{31} & F_{32} \end{pmatrix}, \quad \underline{F}^T = (F_1, F_2, F_3, F_4, F_5, F_6) = (F_{11}, F_{21}, F_{31}, F_{12}, F_{22}, F_{32}). \tag{3.6.17}$$

In the case of $\nabla\underline{\phi}_i$ we define $\bar{\nabla}\underline{\phi}_i$ to mean

$$\left(\bar{\nabla}\underline{\phi}_i\right)^T = \left(\frac{\partial\underline{\phi}_i}{\partial x_1}^T, \frac{\partial\underline{\phi}_i}{\partial x_2}^T\right). \tag{3.6.18}$$

The double dot product operation can be expressed as the product of 6×1 vectors as

$$\frac{\partial W}{\partial \mathbf{F}} : \nabla\underline{\phi}_i = \left(\frac{\partial W}{\partial \underline{F}}\right)^T \bar{\nabla}\underline{\phi}_i = \sum_{r=1}^{6} \frac{\partial W}{\partial F_r} \left(\bar{\nabla}\underline{\phi}_i\right)_r \quad i = 1, 2, \ldots, 9. \tag{3.6.19}$$

Now with $\mathbf{F} = \mathbf{F}(\underline{u})$ the partial derivative with respect to a component of $\underline{u}$ gives, in 3×2 terminology,

$$\frac{\partial \underline{F}}{\partial u_j} = \nabla\underline{\phi}_j, \quad j = 1, 2, \ldots, 9. \tag{3.6.20}$$

Hence if we partially differentiate (3.6.16) with respect to u_j and use the chain rule we have

$$\frac{\partial}{\partial u_j}\left(\frac{\partial W}{\partial \mathbf{F}} : \nabla\underline{\phi}_i\right) = \sum_{r=1}^{6}\sum_{s=1}^{6} \frac{\partial^2 W}{\partial F_s \partial F_r} \left(\bar{\nabla}\underline{\phi}_i\right)_r \left(\bar{\nabla}\underline{\phi}_j\right)_s, \quad i, j = 1, 2, \ldots, 9. \tag{3.6.21}$$

If we swap i and j we have the same quantity and thus we have symmetry in the contribution of $a_1(\cdot;\cdot)$ to the element Jacobian matrix.

The symmetry of the contribution from a_2

The integrand in the a_2 term is

$$\underline{\phi}_i \cdot \left(\frac{\partial \underline{w}}{\partial x_1} \times \frac{\partial \underline{w}}{\partial x_2}\right) + \underline{w} \cdot \left(\frac{\partial \underline{\phi}_i}{\partial x_1} \times \frac{\partial \underline{w}}{\partial x_2}\right) + \underline{w} \cdot \left(\frac{\partial \underline{w}}{\partial x_1} \times \frac{\partial \underline{\phi}_i}{\partial x_2}\right). \tag{3.6.22}$$

Now as

$$\underline{w} = \begin{pmatrix} x_1 \\ x_2 \\ 0 \end{pmatrix} + \sum_{j=1}^{9} u_j \underline{\phi}_j(x_1, x_2) \tag{3.6.23}$$

we have

$$\frac{\partial \underline{w}}{\partial u_j} = \underline{\phi}_j. \tag{3.6.24}$$

Thus if we partially differentiate (3.6.22) with respect to u_j then we have

$$\begin{aligned} & \underline{\phi}_i \cdot \left(\frac{\partial \underline{\phi}_j}{\partial x_1} \times \frac{\partial \underline{w}}{\partial x_2} + \frac{\partial \underline{w}}{\partial x_1} \times \frac{\partial \underline{\phi}_j}{\partial x_2} \right) \\ + \quad & \underline{w} \cdot \left(\frac{\partial \underline{\phi}_i}{\partial x_1} \times \frac{\partial \underline{\phi}_j}{\partial x_2} + \frac{\partial \underline{\phi}_j}{\partial x_1} \times \frac{\partial \underline{\phi}_i}{\partial x_2} \right) \\ + \quad & \underline{\phi}_j \cdot \left(\frac{\partial \underline{\phi}_i}{\partial x_1} \times \frac{\partial \underline{w}}{\partial x_2} + \frac{\partial \underline{w}}{\partial x_1} \times \frac{\partial \underline{\phi}_i}{\partial x_2} \right). \end{aligned} \tag{3.6.25}$$

If we swap i and j we have the same quantity and thus we have symmetry in the contribution of $a_2(\cdot;\cdot)$ to the element Jacobian matrix.

3.6.5 The computation of the derivatives of W

In this thesis we use throughout the constitutive model in the form $W = W(\lambda_1, \lambda_2)$ and, as has just been shown, we need the partial derivatives with respect to the components of $\mathbf{F}$ in the expressions. To get the partial derivatives with respect to the components of $\mathbf{F}$ just involves using the chain rule, as we describe next, although care is needed in evaluating expressions when $\lambda_1 = \lambda_2$ and when these two stretch ratios are very close.

In the following we assume that $W = W(\lambda_1, \lambda_2)$ has a representation in the Ogden form, i.e.

$$W = \sum_1^N \frac{\mu_p}{\nu_p}(\lambda_1^{\nu_p} + \lambda_2^{\nu_p} + \lambda_1^{-\nu_p}\lambda_2^{-\nu_p} - 3), \quad \frac{\mu_p}{\nu_p} > 0. \tag{3.6.26}$$

For shorthand in the expressions we now let

$$W_1 = \frac{\partial W}{\partial \lambda_1} \quad \text{and} \quad W_2 = \frac{\partial W}{\partial \lambda_2}. \tag{3.6.27}$$

and for the second partial derivatives we let

$$W_{11} = \frac{\partial^2 W}{\partial \lambda_1^2}, \quad W_{12} = \frac{\partial^2 W}{\partial \lambda_1 \partial \lambda_2}, \quad W_{22} = \frac{\partial^2 W}{\partial \lambda_2^2}. \tag{3.6.28}$$

By using the chain rule we get

$$\frac{\partial W}{\partial F_s} = W_1 \frac{\partial \lambda_1}{\partial F_s} + W_2 \frac{\partial \lambda_2}{\partial F_s}. \tag{3.6.29}$$

For the second partial derivatives we have

$$\begin{aligned}\frac{\partial^2 W}{\partial F_r \partial F_s} &= W_{11}\frac{\partial \lambda_1}{\partial F_r}\frac{\partial \lambda_1}{\partial F_s} + W_{22}\frac{\partial \lambda_2}{\partial F_r}\frac{\partial \lambda_2}{\partial F_s} + W_{12}\left(\frac{\partial \lambda_1}{\partial F_r}\frac{\partial \lambda_1}{\partial F_s} + \frac{\partial \lambda_2}{\partial F_r}\frac{\partial \lambda_2}{\partial F_s}\right) \\ &\quad + W_1\frac{\partial^2 \lambda_1}{\partial F_r \partial F_s} + W_2\frac{\partial^2 \lambda_2}{\partial F_r \partial F_s}. \end{aligned} \tag{3.6.30}$$

Although we can use (3.6.29) and (3.6.30) for most of the computations we need to adjust when $\lambda_1 = \lambda_2$ as the second derivatives of λ_1 and λ_2 with respect to the components of $\mathbf{F}$ tend to ∞ as $\lambda_1 \to \lambda_2$ as we explain in a moment. The limit of the entire right hand expression is however finite as $\lambda_1 \to \lambda_2$ and we can overcome this difficulty by re-writing as follows. For the last 2 terms in (3.6.30) we write

$$W_1\frac{\partial^2 \lambda_1}{\partial F_r \partial F_s} + W_2\frac{\partial^2 \lambda_2}{\partial F_r \partial F_s} = \left(\frac{W_1 - W_2}{\lambda_1 - \lambda_2}\right)(\lambda_1 - \lambda_2)\frac{\partial^2 \lambda_1}{\partial F_r \partial F_s} + W_2\frac{\partial^2}{\partial F_r \partial F_s}(\lambda_1 + \lambda_2). \tag{3.6.31}$$

Each of the terms

$$\frac{W_1 - W_2}{\lambda_1 - \lambda_2}, \quad (\lambda_1 - \lambda_2)\frac{\partial^2 \lambda_1}{\partial F_r \partial F_s} \quad \text{and} \quad \frac{\partial^2}{\partial F_r \partial F_s}(\lambda_1 + \lambda_2) \tag{3.6.32}$$

have finite limits as $\lambda_1 \to \lambda_2$.

The term $(W_1 - W_2)/(\lambda_1 - \lambda_2)$

With W given in (3.6.26) we have

$$\frac{W_1 - W_2}{\lambda_1 - \lambda_2} = \sum_{p=1}^{N} \mu_p \left(\frac{\lambda_1^{\nu_p} - \lambda_2^{\nu_p}}{\lambda_1 - \lambda_2}\right) \to \sum_{p=1}^{N} \mu_p \nu_p \lambda_1^{\nu_p - 1} \quad \text{as } \lambda_1 \to \lambda_2. \tag{3.6.33}$$

If we need to compute this accurately when $\lambda_1 \neq \lambda_2$ but with λ_1 and λ_2 being very close then we have

$$\lambda_1^{\nu_p} - \lambda_2^{\nu_p} = \int_{\lambda_2}^{\lambda_1} \nu_p x^{\nu_p - 1}\, dx, \tag{3.6.34}$$

by letting $x = \lambda_2 + t(\lambda_1 - \lambda_2)$, $0 \leq t \leq 1$, we get the following

$$\lambda_1^{\nu_p} - \lambda_2^{\nu_p} = (\lambda_1 - \lambda_2)\nu_p \int_0^1 (\lambda_2 + t(\lambda_1 - \lambda_2))^{\nu_p - 1}\, dt. \tag{3.6.35}$$

Therefore we end up we the following expression

$$\frac{\lambda_1^{\nu_p} - \lambda_2^{\nu_p}}{\lambda_1 - \lambda_2} = \nu_p \int_0^1 (\lambda_2 + t(\lambda_1 - \lambda_2))^{\nu_p - 1}\, dt \tag{3.6.36}$$

and we can accurately approximate this using Gauss Legendre quadrature.

The first partial derivatives of λ_1, λ_2 and $\lambda_1 + \lambda_2$

λ_1^2 and λ_2^2 are the eigenvalues of $\mathbf{C}$ and we throughout we choose to take $\lambda_1 \geq \lambda_2$. The characteristic equation of $\mathbf{C}$ is

$$\lambda^2 - (c_{11} + c_{22})\lambda + (c_{11}c_{22} - c_{12}^2) = 0. \tag{3.6.37}$$

Let $\Delta \geq 0$ be such that

$$\Delta^2 = (c_{11} + c_{22})^2 - 4(c_{11}c_{22} - c_{12}^2) = (c_{11} - c_{22})^2 + 4c_{12}^2. \tag{3.6.38}$$

Thus, we get the following expressions for λ_1^2 and λ_2^2

$$\lambda_1^2 = \frac{c_{11} + c_{22} + \Delta}{2} \quad \text{and} \quad \lambda_2^2 = \frac{c_{11} + c_{22} - \Delta}{2}. \tag{3.6.39}$$

Now, for the first partial derivatives of λ_1 and λ_2, by using the chain rule, we get

$$\frac{\partial \lambda_i}{\partial F_s} = \frac{\partial \lambda_i}{\partial c_{11}}\frac{\partial c_{11}}{\partial F_s} + \frac{\partial \lambda_i}{\partial c_{22}}\frac{\partial c_{22}}{\partial F_s} + \frac{\partial \lambda_i}{\partial c_{12}}\frac{\partial c_{12}}{\partial F_s}, \quad \text{for } i = 1, 2 \quad \text{and } s = 1, \ldots, 6. \tag{3.6.40}$$

Since $\mathbf{C} = \mathbf{F}^T\mathbf{F}$,

$$\begin{pmatrix} c_{11} & c_{12} \\ c_{12} & c_{22} \end{pmatrix} = \begin{pmatrix} F_1 & F_2 & F_3 \\ F_4 & F_5 & F_6 \end{pmatrix} \begin{pmatrix} F_1 & F_4 \\ F_2 & F_5 \\ F_3 & F_6 \end{pmatrix} = \begin{pmatrix} F_1^2 + F_2^2 + F_3^2 & F_1F_4 + F_2F_5 + F_3F_6 \\ F_1F_4 + F_2F_5 + F_3F_6 & F_4^2 + F_5^2 + F_6^2 \end{pmatrix}. \tag{3.6.41}$$

By using the above expression, we get the following partial derivatives with respect to $\underline{F}$.

$$\left(\frac{\partial c_{11}}{\partial \underline{F}}\right)^T = 2(F_1, F_2, F_3, 0, 0, 0), \tag{3.6.42}$$

$$\left(\frac{\partial c_{22}}{\partial \underline{F}}\right)^T = 2(0, 0, 0, F_4, F_5, F_6), \tag{3.6.43}$$

$$\left(\frac{\partial c_{12}}{\partial \underline{F}}\right)^T = 2(F_4, F_5, F_6, F_1, F_2, F_3) \tag{3.6.44}$$

Since

$$2\Delta\frac{\partial \Delta}{\partial c_{11}} = 2(c_{11} - c_{22}) \quad \text{we get} \quad \frac{\partial \Delta}{\partial c_{11}} = \frac{c_{11} - c_{22}}{\Delta}. \tag{3.6.45}$$

Similarly

$$\frac{\partial \Delta}{\partial c_{22}} = -\left(\frac{c_{11} - c_{22}}{\Delta}\right) \quad \text{and} \quad \frac{\partial \Delta}{\partial c_{12}} = \frac{4c_{12}}{\Delta} \tag{3.6.46}$$

The case $\lambda_1 = \lambda_2$ is when $\Delta = 0$ which is if and only if $c_{11} = c_{22}$ and $c_{12} = 0$. Hence we cannot evaluate in the form just given but it is sufficient to take the following

$$\frac{\partial \Delta}{\partial c_{11}} = 1, \quad \frac{\partial \Delta}{\partial c_{22}} = -1 \quad \text{and} \quad \frac{\partial \Delta}{\partial c_{12}} = 0. \tag{3.6.47}$$

Then, by using (3.6.39) we get the following partial derivatives with respect to c_{11}

$$2\lambda_1 \frac{\partial \lambda_1}{\partial c_{11}} = \frac{1}{2}\left(1 + \frac{c_{11} - c_{22}}{\Delta}\right), \quad 2\lambda_2 \frac{\partial \lambda_2}{\partial c_{11}} = \frac{1}{2}\left(1 - \frac{c_{11} - c_{22}}{\Delta}\right), \tag{3.6.48}$$

Similarly we get the following partial derivatives with respect to c_{22}

$$2\lambda_1 \frac{\partial \lambda_1}{\partial c_{22}} = \frac{1}{2}\left(1 - \frac{c_{11} - c_{22}}{\Delta}\right), \quad 2\lambda_2 \frac{\partial \lambda_2}{\partial c_{22}} = \frac{1}{2}\left(1 + \frac{c_{11} - c_{22}}{\Delta}\right), \tag{3.6.49}$$

For the partial derivatives with respect to c_{12} we have

$$2\lambda_1 \frac{\partial \lambda_1}{\partial c_{12}} = \frac{4c_{12}}{\Delta}, \quad 2\lambda_2 \frac{\partial \lambda_2}{\partial c_{12}} = -\frac{4c_{12}}{\Delta}. \tag{3.6.50}$$

To be consistent with the above when considering the partial derivatives of Δ when $\lambda_1 = \lambda_2$ we take

$$2\lambda_1 \frac{\partial \lambda_1}{\partial c_{11}} = 1, \quad 2\lambda_2 \frac{\partial \lambda_2}{\partial c_{22}} = 1, \tag{3.6.51}$$

with all the other partial derivatives with respect to the components of $\mathbf{C}$ being 0, when we use the values in (3.6.40).

In the case of the derivatives of $a = \lambda_1 + \lambda_2$ the situation is a bit more straightforward as

$$a^2 = (\lambda_1 + \lambda_2)^2 = \lambda_1^2 + \lambda_2^2 + 2\lambda_1\lambda_2 \tag{3.6.52}$$

$$= c_{11} + c_{22} + 2\sqrt{c_{11}c_{22} - c_{12}^2}. \tag{3.6.53}$$

The first partial derivatives satisfy

$$2a\frac{\partial a}{\partial c_{11}} = 1 + \frac{c_{22}}{\sqrt{c_{11}c_{22} - c_{12}^2}}, \tag{3.6.54}$$

$$2a\frac{\partial a}{\partial c_{22}} = 1 + \frac{c_{11}}{\sqrt{c_{11}c_{22} - c_{12}^2}}, \tag{3.6.55}$$

$$2a\frac{\partial a}{\partial c_{12}} = -2\frac{c_{12}}{\sqrt{c_{11}c_{22} - c_{12}^2}}. \tag{3.6.56}$$

We then get the partial derivatives with respect to $\underline{F}$ by the chain rule, i.e.

$$\frac{\partial a}{\partial F_s} = \frac{\partial a}{\partial c_{11}}\frac{\partial c_{11}}{\partial F_s} + \frac{\partial a}{\partial c_{22}}\frac{\partial c_{22}}{\partial F_s} + \frac{\partial a}{\partial c_{12}}\frac{\partial c_{12}}{\partial F_s}. \tag{3.6.57}$$

The second partial derivatives of λ_1, λ_2 and $\lambda_1 + \lambda_2$

We do not give every detail here and restrict to giving the main steps involved to get something which we can evaluate to get the values needed in a computer program. If g denotes either of λ_1, λ_2 or $\lambda_1 + \lambda_2$ then by partially differentiating with respect to F_r a relation of the type (3.6.40) the product rule and chain rule gives

$$\begin{aligned}\frac{\partial^2 g}{\partial F_r \partial F_s} &= \left(\frac{\partial g}{\partial c_{11}^2}\frac{\partial c_{11}}{\partial F_r} + \frac{\partial^2 g}{\partial c_{11}\partial c_{22}}\frac{\partial c_{22}}{\partial F_r} + \frac{\partial^2 g}{\partial c_{11}\partial c_{12}}\frac{\partial c_{12}}{\partial F_r}\right)\frac{\partial c_{11}}{\partial F_s} + \frac{\partial g}{\partial c_{11}}\frac{\partial^2 c_{11}}{\partial F_r \partial F_s} \\ &+ \left(\frac{\partial^2 g}{\partial c_{22}\partial c_{11}}\frac{\partial c_{11}}{\partial F_r} + \frac{\partial^2 g}{\partial c_{22}^2}\frac{\partial c_{22}}{\partial F_r} + \frac{\partial^2 g}{\partial c_{22}\partial c_{12}}\frac{\partial c_{12}}{\partial F_r}\right)\frac{\partial c_{22}}{\partial F_s} + \frac{\partial g}{\partial c_{22}}\frac{\partial^2 c_{22}}{\partial F_r \partial F_s} \\ &+ \left(\frac{\partial^2 g}{\partial c_{12}\partial c_{11}}\frac{\partial c_{11}}{\partial F_r} + \frac{\partial^2 g}{\partial c_{12}\partial c_{22}}\frac{\partial c_{22}}{\partial F_r} + \frac{\partial^2 g}{\partial c_{12}^2}\frac{\partial c_{12}}{\partial F_r}\right)\frac{\partial c_{12}}{\partial F_s} + \frac{\partial g}{\partial c_{12}}\frac{\partial^2 c_{12}}{\partial F_r \partial F_s}.\end{aligned} \tag{3.6.58}$$

At a stage when these are needed the first derivative terms should have already been obtained and the extra term needed to be able use this are all the second partial derivatives of g with respect to c_{11}, c_{22} and c_{12} and we need all the second partial derivatives of c_{11}, c_{22} and c_{12} with respect all the components of $\mathbf{F}$. The last part can be obtained from (3.6.42)–(3.6.44) and the values are 2 or 0 depending on which term is being considered. The second partial derivatives of g with respect to c_{11}, c_{22} and c_{12} requires partially differentiating the appropriate relation in (3.6.48)–(3.6.50) and (3.6.54)–(3.6.56). For example, partially

differentiating the first relation in (3.6.48) with respect to c_{11} gives

$$\begin{aligned} 4\lambda_1\frac{\partial^2\lambda_1}{\partial c_{11}^2}+4\left(\frac{\partial\lambda_1}{\partial c_{11}}\right)^2 &= \frac{1}{\Delta}+(c_{11}-c_{22})\frac{\partial}{\partial c_{11}}\left(\frac{1}{\Delta}\right) \\ &= \frac{1}{\Delta}\left(1-\left(\frac{c_{11}-c_{22}}{\Delta}\right)^2\right). \end{aligned} \tag{3.6.59}$$

This is enough to indicate why the second partial derivative tends to ∞ in most ways in which c_{11}, c_{22} and c_{12} can vary with $\Delta\to 0$ and it also indicates that

$$\Delta\frac{\partial^2\lambda_1}{\partial c_{11}^2} \quad \text{and} \quad (\lambda_1-\lambda_2)\frac{\partial^2\lambda_1}{\partial c_{11}^2}=\left(\frac{\Delta}{\lambda_1+\lambda_2}\right)\frac{\partial^2\lambda_1}{\partial c_{11}^2} \tag{3.6.60}$$

each have finite limit as $\Delta\to 0$.

3.6.6 Comments about the existence of the solution

In subsection 3.6.1 we discussed some aspects of starting from a prestretched state and then in subsection 3.6.2 we described the procedure for attempting to get the solution at a sequence of increasing pressures $0=P_0<P_1<\cdots$. It depends on the shape of Ω and the form of the strain energy function W as to whether or not we can only solve the problem in this way for a limited range of pressures as the set-up can be such that we reach the situation where we have a solution $\underline{u}$ at pressure P_k with the Jacobian matrix $J_A(\underline{u};P_k)$ being singular. When this is the case we cannot use this solution to attempt to get the solution at a nearby pressure P_{k+1} and indeed there may not be a solution at a pressure $P_{k+1}>P_k$ or at least there may not be a solution which is close to $\underline{u}$. This is a feature of nonlinear problems of this kind and the critical values of $\underline{u}$ and P_k where this first occurs is typically a limit point and to proceed further needs some path following technique which is not considered in this thesis. For the quasi-static problems we restrict in this thesis to deformations for which J_A is positive definite and recall that in subsection 3.6.1 it was claimed that this is the case when we start with a prestretch. The purpose of this subsection is to give some details as to why this is the case and we give some justification for the existence of the solution for pressures which are not too large.

Let $\underline{u}_0\neq\underline{0}$ denote the prestretched state at pressure $P_0=0$ and let

$$A'(\underline{u};\underline{\alpha},\underline{v},P)=\frac{\mathrm{d}}{\mathrm{d}s}A'(\underline{u}+s\underline{\alpha};\underline{v},P)\bigg|_{s=0} \tag{3.6.61}$$

denote the first Gâteaux derivative of A. (We use Gâteaux derivatives much more in the next chapter.) If we let P denote a very small pressure and we let $\underline{d} = \underline{u} - \underline{u}_0$ be the change in the solution then

$$A'(\underline{u}_0 + \underline{d}; \underline{v}, P) = A(\underline{u}_0; \underline{v}, P) + A'(\underline{u}_0; \underline{d}, \underline{v}, P) + (\text{smaller terms of magnitude } \|\underline{d}\|^2) \tag{3.6.62}$$

and what we refer to as the linearised problem about $\underline{u}_0$ is to find $\underline{d}$ such that

$$A'(\underline{u}_0; \underline{d}, \underline{v}, P) = -A(\underline{u}_0; \underline{v}, P), \quad \forall \underline{v} = (v_i),\ v_i \in H_0^1(\Omega). \tag{3.6.63}$$

This is a linear problem for $\underline{d}$ and the finite element discretized version is the first step of Newton's method for the nonlinear problem. We consider next some properties of this linear problem to justify why it has a solution. For the right hand side in (3.6.63) we have

$$\begin{aligned} -A(\underline{u}_0; \underline{v}, P) &= -a_1(\underline{u}_0; \underline{v}) + P a_2(\underline{u}_0; \underline{v}) \\ &= -h_0 \iint_\Omega \frac{\partial W}{\partial \mathbf{F}} : \nabla \underline{v}\, \mathrm{d}x_1 \mathrm{d}x_2 \\ &\quad + P \iint_\Omega \underline{v} \cdot \left(\frac{\partial \underline{w}_0}{\partial x_1} \times \frac{\partial \underline{w}_0}{\partial x_2} \right) \mathrm{d}x_1 \mathrm{d}x_2, \end{aligned} \tag{3.6.64}$$

where $\underline{w}_0 = \underline{x} + \underline{u}_0$ is the prestretched state and where $\partial W/\partial \mathbf{F}$ is evaluated at $\mathbf{F}_0 = \mathbf{I} + \nabla \underline{u}_0$. With the first Gâteaux derivatives of a_1 and a_2 defined in a similar way to that of Gâteaux derivative of A we have for the left hand side of (3.6.63) that

$$A'(\underline{u}_0; \underline{d}, \underline{v}, P) = a_1'(\underline{u}_0; \underline{d}, \underline{v}) - P a_2'(\underline{u}_0; \underline{d}, \underline{v}) \tag{3.6.65}$$

where

$$a_1'(\underline{u}_0; \underline{d}, \underline{v}) = h_0 \iint_\Omega \sum_{r=1}^{6} \sum_{s=1}^{6} \frac{\partial^2 W}{\partial F_s \partial F_r} (\bar{\nabla} \underline{d})_r (\bar{\nabla} \underline{d})_s \, \mathrm{d}x_1 \mathrm{d}x_2 \tag{3.6.66}$$

and

$$\begin{aligned} a_2'(\underline{u}_0; \underline{d}, \underline{v}) &= \iint_\Omega \underline{v} \cdot \left(\frac{\partial \underline{d}}{\partial x_1} \times \frac{\partial \underline{w}_0}{\partial x_2} + \frac{\partial \underline{w}_0}{\partial x_1} \times \frac{\partial \underline{d}}{\partial x_2} \right) \mathrm{d}x_1 \mathrm{d}x_2 \qquad (3.6.67) \\ &= \frac{1}{3} \iint_\Omega \underline{v} \cdot \left(\frac{\partial \underline{d}}{\partial x_1} \times \frac{\partial \underline{w}_0}{\partial x_2} + \frac{\partial \underline{w}_0}{\partial x_1} \times \frac{\partial \underline{d}}{\partial x_2} \right) \mathrm{d}x_1 \mathrm{d}x_2 \\ &\quad + \frac{1}{3} \iint_\Omega \underline{w}_0 \cdot \left(\frac{\partial \underline{v}}{\partial x_1} \times \frac{\partial \underline{d}}{\partial x_2} + \frac{\partial \underline{d}}{\partial x_1} \times \frac{\partial \underline{v}}{\partial x_2} \right) \mathrm{d}x_1 \mathrm{d}x_2 \\ &\quad + \frac{1}{3} \iint_\Omega \underline{d} \cdot \left(\frac{\partial \underline{v}}{\partial x_1} \times \frac{\partial \underline{w}_0}{\partial x_2} + \frac{\partial \underline{w}_0}{\partial x_1} \times \frac{\partial \underline{v}}{\partial x_2} \right) \mathrm{d}x_1 \mathrm{d}x_2. \qquad (3.6.68) \end{aligned}$$

The expression in the integrand of (3.6.66) corresponds to what is given in (3.6.21) and

the expression in the integrand of the symmetric version in (3.6.67) corresponds to what is given in (3.6.26). To be able to use the Lax-Milgram theorem to show that (3.6.63) has a unique solution when P is sufficiently small we need to verify that the conditions of the theorem hold.

Firstly, the vector $\underline{d} = \underline{d}(x_1, x_2)$ that we seek is $\underline{0}$ on $\partial\Omega$ and thus the function space is $V = H_0^1(\Omega)$ and a norm on this space is such that

$$\|\underline{v}\|_{H_0^1}^2 = \iint_\Omega \left(\frac{\partial v_1}{\partial x_1}\right)^2 + \left(\frac{\partial v_2}{\partial x_1}\right)^2 + \left(\frac{\partial v_3}{\partial x_1}\right)^2 + \left(\frac{\partial v_1}{\partial x_2}\right)^2 + \left(\frac{\partial v_2}{\partial x_2}\right)^2 + \left(\frac{\partial v_3}{\partial x_2}\right)^2 \mathrm{d}x_1\mathrm{d}x_2. \tag{3.6.69}$$

It is useful to also give other similar notation here for the norms of vector valued function defined on Ω.

$$\|\underline{v}\|_{L_2}^2 = \iint_\Omega v_1^2 + v_2^2 + v_3^2 \,\mathrm{d}x_1\mathrm{d}x_2, \tag{3.6.70}$$

$$\|\underline{v}\|_{H^1}^2 = \|\underline{v}\|_{L_2}^2 + \|\underline{v}\|_{H_0^1}^2. \tag{3.6.71}$$

By Friedrich's inequality, see e.g. [10, p.104], the norm $\|.\|_{H_0^1}$ for functions in H_0^1 is equivalent to the norm of $H^1(\Omega)$ and this equivalence of norms means that if we show that the linear functional on the right hand side of (3.6.63) is bounded in one norm then it is also bounded in the other norm. Similarly, if we get a lower bound for $A'(\underline{u}_0; \underline{v}, \underline{v}, P)$ in one of the norms, when P is sufficiently small, then there is also a similar bound in the other norm.

We consider now bounding the linear functional given in (3.6.64). Now recall that the term $\partial W/\partial \mathbf{F}$ is evaluated at the prestretch state and does not vary in Ω and $\partial W/\partial \mathbf{F} : \nabla \underline{v}$ is of the form

$$k_{11}\frac{\partial v_1}{\partial x_1} + k_{21}\frac{\partial v_2}{\partial x_1} + k_{31}\frac{\partial v_3}{\partial x_1} + k_{12}\frac{\partial v_1}{\partial x_2} + k_{22}\frac{\partial v_2}{\partial x_2} + k_{32}\frac{\partial v_3}{\partial x_2}.$$

If we let

$$k_m = \max\{|k_{ij}| : 1 \le i \le 3,\, 1 \le j \le 2\}$$

then by using the triangle inequality and the Cauchy Schwarz inequality

$$\begin{aligned}
\left| h_0 \iint_\Omega \frac{\partial W}{\partial \mathbf{F}} : \nabla \underline{v} \,\mathrm{d}x_1\mathrm{d}x_2 \right| &\le h_0 k_m \iint_\Omega \left|\frac{\partial v_1}{\partial x_1}\right| + \cdots + \left|\frac{\partial v_3}{\partial x_2}\right| \mathrm{d}x_1\mathrm{d}x_2 \\
&\le h_0 k_m \sqrt{(\text{area of } \Omega)} \left(\left\|\frac{\partial v_1}{\partial x_1}\right\|_{L_2} + \cdots + \left\|\frac{\partial v_3}{\partial x_2}\right\|_{L_2} \right) \\
&\le 6 h_0 k_m \sqrt{(\text{area of } \Omega)}\, \|\underline{v}\|_{H_0^1}.
\end{aligned} \tag{3.6.72}$$

For the part in the expression in (3.6.64) involving the cross product we let $\hat{k}_m$ denote the largest of the magnitudes of the entries of the vector $\frac{\partial \underline{w}_0}{\partial x_1} \times \frac{\partial \underline{w}_0}{\partial x_2}$. By the triangle inequality and the Cauchy Schwarz inequality we have

$$\begin{aligned}\left| P \iint_\Omega \underline{v} \cdot \left(\frac{\partial \underline{w}_0}{\partial x_1} \times \frac{\partial \underline{w}_0}{\partial x_2} \right) \mathrm{d}x_1 \mathrm{d}x_2 \right| &\leq P\hat{k}_m \iint_\Omega |v_1| + |v_2| + |v_3| \, \mathrm{d}x_1 \mathrm{d}x_2 \\ &\leq P\hat{k}_m \sqrt{(\text{area of } \Omega)} \left((\|v_1\|_{L_2} + \|v_2\|_{L_2} + \|v_3\|_{L_2} \right) \\ &\leq 3P\hat{k}_m \sqrt{(\text{area of } \Omega)} \, \|\underline{v}\|_{L_2} . \end{aligned} \tag{3.6.73}$$

From (3.6.72) and (3.6.73) we define C_m to be the larger of the constants, i.e. $C_m = \max\left\{6h_0k_m, 3P\hat{k}_k\right\}$ (area of Ω) so that we have

$$|A(\underline{u}_0; \underline{v}, P)| \leq C_m \, \|\underline{v}\|_{H^1} \tag{3.6.74}$$

and we have verified that the right hand side of (3.6.63) defines a bounded linear functional.

We now consider the left hand side term in (3.6.63). With suitable properties for the strain energy density W similar reasoning to what has been given above shows that the bilinear form satisfies a bounded property of the form

$$|A'(\underline{u}_0; \underline{\alpha}, \underline{v}, P)| \leq (\text{const.}) \, \|\underline{\alpha}\|_{H^1} \, \|\underline{v}\|_{H^1} . \tag{3.6.75}$$

We just consider some details to justify that the bilinear form satisfies a coercivity property when the pressure P is sufficiently small. Now from (3.6.65) we have when $\underline{d} = \underline{v}$

$$A'(\underline{u}_0; \underline{v}, \underline{v}, P) = a_1'(\underline{u}_0; \underline{v}, \underline{v}) - P a_2'(\underline{u}_0; \underline{v}, \underline{v}). \tag{3.6.76}$$

For the a_1' term we need properties of the 6×6 matrix $(\partial^2 W / \partial F_r \partial F_s)$. For the strain energy density functions that are considered this matrix is positive definite when the sheet is in tension corresponding to the stretch ratios $\lambda_1 \geq \lambda_2 > 1$ which is what we have when we have a prestretch. We could give fairly lengthy details to verify this but the details are already available in [3]. In [3] it is shown that $(\partial^2 W / \partial F_r \partial F_s)$ when evaluated at a general deformation $(F_1, F_2, F_3, F_4, F_5, F_6)^T$ is positive definite if and only if the corresponding matrix evaluated at a diagonal $\mathbf{F}$ with the same principal values λ_1, λ_2 is positive definite. In this context the diagonal $\mathbf{F}$ in 3×2 form is

$$\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \\ 0 & 0 \end{pmatrix}$$

and in 6×1 vector form corresponds to $(\lambda_1, 0, 0, 0, \lambda_2, 0)^T$. As given in [3] the expression for the 6×6 matrix when evaluated with such an $\mathbf{F}$ is a relatively short expression and is given by

$$\begin{pmatrix} W_{11} & 0 & 0 & 0 & W_{12} & 0 \\ 0 & G & 0 & H & 0 & 0 \\ 0 & 0 & W_1/\lambda_1 & 0 & 0 & 0 \\ 0 & H & 0 & G & 0 & 0 \\ W_{12} & 0 & 0 & 0 & W_{22} & 0 \\ 0 & 0 & 0 & 0 & 0 & W_2/\lambda_2 \end{pmatrix}, \qquad G = \frac{\lambda_1 W_1 - \lambda_2 W_2}{\lambda_1^2 - \lambda_2^2}, \quad H = \frac{\lambda_2 W_1 - \lambda_1 W_2}{\lambda_1^2 - \lambda_2^2}. \tag{3.6.77}$$

When $\lambda_1 \geq \lambda_2 > 1$ the positive definite property is that $W_1 > 0$, $W_2 > 0$ and the following two 2×2 matrices

$$\begin{pmatrix} W_{11} & W_{12} \\ W_{12} & W_{22} \end{pmatrix}, \quad \begin{pmatrix} G & H \\ H & G \end{pmatrix},$$

are positive definite. With a prestretch as described this is constant throughout Ω and if we let $\mu_1 > 0$ denote the smallest eigenvalue then we have

$$a_1'(\underline{u}_0; \underline{v}, \underline{v}) \geq \mu_1 \left\| \underline{v} \right\|_{H_0^1}^2. \tag{3.6.78}$$

This is the coercivity property of the first part on the left hand side of (3.6.63).

When P is sufficiently small the a_2' term cannot change much the lower bound for $A'(\underline{u}_0; \underline{v}, \underline{v}, P)$ provided we can show that the a_2' term is suitably bounded. Now, as given in (3.6.67) and (3.6.68), we have two ways to represent $a_2'(\underline{u}_0; \underline{v}, \underline{v})$ and for the bound we can just take the shorter version (3.6.67). Let

$$\underline{q} = \frac{\partial \underline{v}}{\partial x_1} \times \frac{\partial \underline{w}_0}{\partial x_2} + \frac{\partial \underline{w}_0}{\partial x_1} \times \frac{\partial \underline{v}}{\partial x_2} \tag{3.6.79}$$

and in the following we use $|.|$ for the Euclidean length of a vector. By the triangle inequality for vectors and properties of the cross product we have

$$\begin{aligned} \left| \underline{q} \right| &\leq \left| \frac{\partial \underline{v}}{\partial x_1} \times \frac{\partial \underline{w}_0}{\partial x_2} \right| + \left| \frac{\partial \underline{w}_0}{\partial x_1} \times \frac{\partial \underline{v}}{\partial x_2} \right| \\ &\leq \left| \frac{\partial \underline{v}}{\partial x_1} \right| \left| \frac{\partial \underline{w}_0}{\partial x_2} \right| + \left| \frac{\partial \underline{w}_0}{\partial x_1} \right| \left| \frac{\partial \underline{v}}{\partial x_2} \right|. \end{aligned}$$

Let

$$\tilde{k}_m = \max \left\{ \left| \frac{\partial \underline{w}_0}{\partial x_1} \right|, \left| \frac{\partial \underline{w}_0}{\partial x_2} \right| \right\}.$$

Using this we have

$$\begin{aligned}
\left|\underline{q}\right| &\leq \tilde{k}_m \left(\left| \frac{\partial \underline{v}}{\partial x_1} \right| + \left| \frac{\partial \underline{v}}{\partial x_2} \right| \right), \\
\left|\underline{q}\right|^2 &\leq \tilde{k}_m^2 \left(\left| \frac{\partial \underline{v}}{\partial x_1} \right|^2 + \left| \frac{\partial \underline{v}}{\partial x_2} \right|^2 + 2 \left| \frac{\partial \underline{v}}{\partial x_1} \right| \left| \frac{\partial \underline{v}}{\partial x_2} \right| \right) \\
&\leq 2\tilde{k}_m^2 \left(\left| \frac{\partial \underline{v}}{\partial x_1} \right|^2 + \left| \frac{\partial \underline{v}}{\partial x_2} \right|^2 \right).
\end{aligned}$$

By the Cauchy Schwarz inequality we then have that

$$\begin{aligned}
a_2'(\underline{u}_0; \underline{v}, \underline{v}) &\leq 2\tilde{k}_m \left(\|v_1\|_{L_2} \|q_1\|_{L_2} + \|v_2\|_{L_2} \|q_2\|_{L_2} + \|v_3\|_{L_2} \|q_3\|_{L_2} \right) \\
&\leq 6\tilde{k}_m \|\underline{v}\|_{L_2} \|\underline{v}\|_{H_0^1}.
\end{aligned}$$

As $\|\underline{v}\|_{L_2} \leq \|\underline{v}\|_{H^1}$ and $\|\underline{v}\|_{H_0^1} \leq \|\underline{v}\|_{H^1}$ we have a bound

$$a_2'(\underline{u}_0; \underline{v}, \underline{v}) \leq 6\tilde{k}_m \|\underline{v}\|_{H^1}^2. \tag{3.6.80}$$

From the earlier comment about the equivalence of the norms $\|.\|_{H^1}$ and $\|.\|_{H_0^1}$ for functions in $H_0^1(\Omega) \subset H^1(\Omega)$ it follows from (3.6.78) and (3.6.80) that when $P > 0$ is sufficiently small there exists a constant $c_1 > 0$ such that

$$A'(\underline{u}_0; \underline{v}, \underline{v}, P) \geq c_1 \|\underline{v}\|_{H_0^1}^2. \tag{3.6.81}$$

This completes the details to verify that the conditions of the Lax-Milgram theorem hold for the linear problem in (3.6.63) to have a unique solution in $H_0^1(\Omega)$.

We finish this subsection with comments about the situation when there is no pre-stretch which corresponds to $\underline{u}_0 = \underline{0}$, $\underline{w}_0 = \underline{x}$, $\mathbf{F} = \mathbf{I}$ and $\lambda_1 = \lambda_2 = 1$. When this is the case $W_1 = W_2 = 0$ and all components of the stress $\mathbf{\Pi}$ are 0. We do not have the coercivity property and in particular if $\underline{v} = (0, 0, v_3)^T$ with $v_3 = v_3(x_1, x_2)$ not being identical zero on Ω then $a_1'(\underline{0}; \underline{v}, \underline{v}) = 0$. We have a similar situation with the a_2' term since we have

$$a_2'(\underline{0}; \underline{v}, \underline{v}) = \iint_\Omega v_3 \underline{e}_3 \cdot \left(\frac{\partial v_3}{\partial x_1} \underline{e}_3 \times \underline{e}_2 + \underline{e}_1 \times \frac{\partial v_3}{\partial x_2} \underline{e}_3 \right) \mathrm{d}x_1 \mathrm{d}x_2 = 0. \tag{3.6.82}$$

In the numerical scheme the rows of the Jacobian matrix associated with entries in the $\underline{e}_3$ direction are all zero when we evaluate it at $\underline{u} = \underline{0}$ and thus it is a singular matrix. Hence, as already mentioned, we cannot start a Newton iteration with a vector corresponding to $\underline{u} = \underline{0}$.

4. THE DUAL PROBLEM FOR ERROR ESTIMATION IN A QOI

4.1 Introduction and some of the notation

This chapter is mainly concerned with describing, in an abstract way, how to represent the error in an approximation to a QoI when the problem which determines the primary unknown is a weak form description of a system of nonlinear PDEs. The first of the specific weak forms that is considered is the membrane problem given in (3.4.15)–(3.4.17), once a strain energy function has been specified, and this is considered further in the last section of this chapter. The other two weak forms that are considered are given in chapter 6 when the axisymmetric membrane problem is considered in the quasi-static case and in the dynamic cases. The estimation of the error in a QoI for the problems in chapter 6 and considered in chapter 7.

In the abstract description, we show that the difference between the QoI we wish to compute and the approximate value we actually compute can be represented in terms of a function which satisfies a related dual problem and we cannot in general solve the dual problem exactly. If we refer to the dual problem which gives the exact representation as the exact dual problem then we discuss different ways in which it might be approximated which gives us different possible computable schemes.

To describe the main problem and the QoI we use the following notation.

$$
\begin{aligned}
V &= \text{infinite dimensional Hilbert space of functions defined on } \Omega,\\
V_h &= \text{finite element space involving piecewise polynomials where } V_h \subset V,\\
\underline{U} &= \text{exact solution in } V \text{ to the weak problem},\\
\underline{U}_h &= \text{finite element approximation in } V_h \ ,\\
\underline{e}_h &= \underline{U} - \underline{U}_h = \text{error in the approximation of } \underline{U}_h \text{ to } \underline{U},\\
J &= \text{QoI functional},\\
J(\underline{U}) &= \text{the quantity we wish to compute},\\
J(\underline{U}_h) &= \text{our estimate of the QoI},\\
A(\cdot;\cdot) &= \text{semi-linear form on } V \times V \text{ in the problem defining } \underline{U}, \text{ (see comment below)},\\
F(\cdot) &= \text{linear functional on } V \text{ in the problem defining } \underline{U} \text{ (see comment below)}.
\end{aligned}
$$

Notes

(i) h denotes a mesh size and when convergence is mentioned we mean as $h \to 0$.

(ii) Ω denotes a generic domain in both space and time region which we define later appropriately.

(iii) The semi-linear form $A(\cdot;\cdot)$ is linear in terms after the semi-colon and in all our cases it is nonlinear in terms before the semi-colon. The argument before and after the semi-colon are functions in the Hilbert space V being considered. Similarly the the argument of the linear functional $F(\cdot)$ is also for a function in the Hilbert space V being considered. Thus in particular if $\underline{U}, \underline{v}_1, \underline{v}_2 \in V$ and $\alpha_1, \alpha_2 \in \mathbb{R}$ then

$$
\begin{aligned}
A(\underline{U}; \alpha_1\underline{v}_1 + \alpha_2\underline{v}_2) &= \alpha_1 A(\underline{U}; \underline{v}_1) + \alpha_2 A(\underline{U}; \underline{v}_2), &(4.1.1)\\
F(\alpha_1\underline{v}_1 + \alpha_2\underline{v}_2) &= \alpha_1 F(\underline{v}_1) + \alpha_2 F(\underline{v}_2). &(4.1.2)
\end{aligned}
$$

In cetain places in what follows we will need a degree of smoothness of the semi-linear form $A(\cdot;\cdot)$ involving Gâteaux derivatives be bounded and we define a Gâteaux derivative in section 4.2. Another property that we require in some places later in the chapter is that the first Gâteaux derivative satisfies a coercive property and what we explain what this means in this context when the property is needed.

We not not consider details sufficient to guarantee the existence of a solution in an abstract setting and restrict to introducing the problems which we assume have solutions.

We assume that we have a Hilbert space V an appropriate semi-linear form $A(\cdot;\cdot)$ and a linear functional $F(\cdots)$ such that there exists a unique solution $\underline{U} \in V$ satisfying

$$A(\underline{U};\underline{\psi}) = F(\underline{\psi}), \quad \forall \underline{\psi} \in V. \tag{4.1.3}$$

Similarly we assume that there exists a unique solution $\underline{U}_h \in V_h$ satisfying

$$A(\underline{U}_h;\underline{\psi}) = F(\underline{\psi}), \quad \forall \underline{\psi} \in V_h. \tag{4.1.4}$$

We refer to the solution $\underline{U}$ satisfying (4.1.3) as the exact solution and we refer to $\underline{U}_h$ as the finite element approximation of $\underline{U}$. The aim of a computation is to compute an estimate $J(\underline{U}_h)$ of $J(\underline{U})$ of sufficient accuracy and in the next section we show a way of representing the error $J(\underline{U}) - J(\underline{U}_h)$.
Furthermore, when $V_h \subset V$ the subtraction of (4.1.4) from (4.1.3) gives

$$A(\underline{U};\underline{\psi}) - A(\underline{U}_h;\underline{\psi}) = 0 \quad \forall \underline{\psi} \in V_h. \tag{4.1.5}$$

which is similar to the Galerkin orthogonality result in the linear case.

As a final point here, it is worth noting that for quasi-static membrane inflation problems considered in this thesis the term $F(\underline{\psi}) = 0$ as we have already given in (3.6.1). This is not really any significant simplification or special case but is just as a consequence of the pressure loading term being part of the $A(\cdot;\cdot)$ term. As two of the three weak problems being considered in this thesis are of this type it is worth indicating here how the problems will appear in these cases. Specifically, in these case the weak forms will appear as follows. We need to find the solution $\underline{U} \in V$ that satisfies

$$A(\underline{U};\underline{\psi}) = \underline{0}, \quad \forall \underline{\psi} \in V \tag{4.1.6}$$

Similarly when $F(\underline{\psi}) = 0$, the finite element solution $\underline{U}_h \in V_h$ satisfies

$$A(\underline{U}_h;\underline{\psi}) = \underline{0}, \quad \forall \underline{\psi} \in V_h. \tag{4.1.7}$$

4.2 A representation of $J(\underline{U}) - J(\underline{U}_h)$ and a dual solution $\underline{\psi}$

Many of the details presented can be found in [17] and [28] and see also [8].

In the case of a differentiable function $g : \mathbb{R} \to \mathbb{R}$, the fundamental theorem of calculus

gives us a representation for the difference between two function values as

$$g(c) - g(b) = \int_b^c g'(s)\, ds.$$

We can extend this idea to functionals with the derivative of a function replaced by the Gâteaux derivative of a functional in the direction of another function which we next define. Let $\underline{\psi} \in V$ be fixed. The first Gâteaux derivatives of $J(\cdot)$ and $A(\cdot;\underline{\psi})$ in the direction of $\underline{\alpha} \in V$ are defined as follows.

$$J'\left(\underline{u};\underline{\alpha}\right) := \frac{d}{ds} J\left(\underline{u} + s\underline{\alpha}\right)\Big|_{s=0} \quad \text{and} \quad A'\left(\underline{u};\underline{\alpha},\underline{\psi}\right) := \frac{d}{ds} A\left(\underline{u} + s\underline{\alpha};\underline{\psi}\right)\Big|_{s=0}. \tag{4.2.1}$$

Now if we take the direction $\underline{\alpha} = \underline{e}_h = \underline{U} - \underline{U}_h$ and we let

$$g(s) = J(\underline{U}_h + s\underline{e}_h), \quad 0 \le s \le 1, \tag{4.2.2}$$

then

$$g'(s) = J'(\underline{U}_h + s\underline{e}_h;\underline{e}_h), \quad 0 \le s \le 1, \tag{4.2.3}$$

and

$$J(\underline{U}) - J(\underline{U}_h) = g(1) - g(0) = \int_0^1 g'(s)\, ds = \int_0^1 J'(\underline{U}_h + s\underline{e}_h;\underline{e}_h)\, ds. \tag{4.2.4}$$

We can similarly write

$$A(\underline{U};\underline{\psi}) - A(\underline{U}_h;\underline{\psi}) = \int_0^1 A'(\underline{U}_h + s\underline{e}_h;\underline{e}_h,\underline{\psi})\, ds \tag{4.2.5}$$

and since $\underline{U}$ satisfies (4.1.3) we have

$$F(\underline{\psi}) - A(\underline{U}_h;\underline{\psi}) = \int_0^1 A'(\underline{U}_h + s\underline{e}_h;\underline{e}_h,\underline{\psi})\, ds. \tag{4.2.6}$$

As we do not know the error direction $\underline{e}_h$ we consider all possible directions $\underline{\alpha}$, which will include $\underline{e}_h$, and we now let $\underline{\psi} \in V$ be such that

$$\int_0^1 A'\left(\underline{U}_h + s\underline{e}_h;\underline{\alpha},\underline{\psi}\right) ds = \int_0^1 J'\left(\underline{U}_h + s\underline{e}_h;\underline{\alpha}\right) ds \quad \forall \underline{\alpha} \in V. \tag{4.2.7}$$

This is a linear problem and it is what we refer to as the exact dual problem partly because the first argument in the terms $A'(\cdot;\cdot,\cdot)$ and $J'(\cdot;\cdot)$ involve the exact solution $\underline{U}$. We assume that the set-up is such that there exists a unique solution to this dual problem and, without going into details, sufficient conditions for this would be a coercive

and bounded property of $A'(\underline{U}_h + s\underline{e}_h; \cdot, \cdot)$ for all $0 \leq s \leq 1$, a bounded property of $J'(\underline{u}_h + s\underline{e}_h)$ for all $0 \leq s \leq 1$ and continuity of both terms with respect to s. Accepting that there is such a $\underline{\phi}$ it follows that by using (4.2.5) and (4.2.4) the representation of the error in the approximation of the functional is given by

$$J(\underline{U}) - J(\underline{U}_h) = F(\underline{\psi}) - A(\underline{U}_h; \underline{\psi}). \tag{4.2.8}$$

The solution to the exact dual problem hence gives us an exact representation of the error.

There are other ways we can write (4.2.8) as a consequence of the function $\underline{U}_h$ satisfying (4.1.4), e.g.

$$J(\underline{U}) - J(\underline{U}_h) = F(\underline{\psi} - \underline{v}_h) - A(\underline{U}_h; \underline{\psi} - \underline{v}_h) \quad \forall \underline{v}_h \in V_h. \tag{4.2.9}$$

Now for the problems considered in the thesis, and these are covered in section 4.6 and chapter 7, the expression for $F(\cdot)$ and $A(\cdot;\cdot)$ involve integrating over a domain Ω and in the finite element method Ω is partitioned into elements $\Omega_1, \ldots, \Omega_{ne}$. If we let $F(\cdot)_k$ and $A(\cdot;\cdot)_k$ denote the versions of $F(\cdot)$ and $A(\cdot;\cdot)$ when we just integrate over the element Ω_k then we have

$$J(\underline{U}) - J(\underline{U}_h) = \sum_{k=1}^{ne} F(\underline{\psi} - \underline{v}_h)_k - \sum_{k=1}^{ne} A(\underline{U}_h; \underline{\psi} - \underline{v}_h)_k \quad \forall \underline{v}_h \in V_h. \tag{4.2.10}$$

If we take $\underline{v}_h = \underline{\psi}_I$ to be the interpolant of $\underline{\psi}$ in V_h, or we take some other projection of $\underline{\psi}$ in the function space, then quantities such as

$$F(\underline{\psi} - \underline{\psi}_I)_k - A(\underline{U}_h; \underline{\psi} - \underline{\psi}_I)_k, \quad k = 1, \ldots, ne \tag{4.2.11}$$

are taken as the element contributions to the value $J(\underline{U}) - J(\underline{U}_h)$ and when we can adequately approximate $\underline{\psi}$ the corresponding quantities may be used to drive an adaptive scheme.

4.3 The rate at which $J(\underline{U}) - J(\underline{U}_h)$ tends to 0 as $h \to 0$

With a representation of the error we can now make some theoretical comments about how rapidly it decreases as a mesh is refined.

From the representation and the Galerkin orthogonality type property given in (4.1.5)

we have that for all $\underline{\psi}_h \in V_h$

$$\begin{aligned} J(\underline{U}) - J(\underline{U}_h) &= F(\underline{\psi}) - A(\underline{U}_h; \underline{\psi}) \\ &= A(\underline{U}; \underline{\psi}) - A(\underline{U}_h; \underline{\psi}) \\ &= A(\underline{U}; \underline{\psi} - \underline{\psi}_h) - A(\underline{U}_h; \underline{\psi} - \underline{\psi}_h) \\ &= \int_0^1 A'(\underline{U}_h + s\underline{e}_h; \underline{e}_h, \underline{\psi} - \underline{\psi}_h)\, \mathrm{d}s. \end{aligned} \tag{4.3.1}$$

The expression $A'(\underline{U}_h + s\underline{e}_h; \underline{e}_h, \underline{\psi} - \underline{\psi}_h)$ is linear in the terms $\underline{e}_h$ and $\underline{\psi} - \underline{\psi}_h$ which appear after the semi-colon and thus we have a product of the components of such terms. Now it depends on the expression for $A(\cdot;\cdot)$ which generates $A'(\cdot;\cdot,\cdot)$ what can be said and in the case that $A'(\underline{U}_h + s\underline{e}_h; \cdot, \cdot)$ is bounded for all $s \in [0,1]$ and the terms to the right of the semi-colon involve function values and first derivative values this leads to

$$|J(\underline{U}) - J(\underline{U}_h)| \leq (\text{const}) \, \|\underline{e}_h\|_{H^1(\Omega)} \left\| \underline{\psi} - \underline{\psi}_h \right\|_{H^1(\Omega)} \quad \forall \underline{\psi}_h \in V_h. \tag{4.3.2}$$

When piecewise polynomials of degree p are used in the finite element approximation the best that we can have for the error is that $\|\underline{e}_h\|_{H^1}$ is $\mathcal{O}(h^p)$ as discussed in the preliminary chapter, see also [10, p.102]. Also, when $A'(\cdot;\cdot,\cdot)$ and $J'(\cdot)$ are such that $\underline{\psi}$ is sufficiently smooth we know that there exists $\underline{\psi}_h \in V_h$ such that $\left\| \underline{\psi} - \underline{\psi}_h \right\|_{H^1(\Omega)}$ is $\mathcal{O}(h^p)$ (a suitable interpolant has this property, see e.g. [10, p.109]). As we have the product of two terms which are each $\mathcal{O}(h^p)$ we deduce that when all the conditions are met

$$J(\underline{U}) - J(\underline{U}_h) = \mathcal{O}(h^{2p}). \tag{4.3.3}$$

If $\underline{\psi}$ is not sufficiently smooth then the rate of convergence is less which is the case, for example, if the functional $J(\cdot)$ involves pointwise values of the components of $\underline{U}$. More details of the above result can be found in [23].

4.4 Possible dual problems to solve

From the representation of the error given in (4.2.8) it follows that we need an approximation to $\underline{\psi}$ which is from a different space than V_h as otherwise our estimate of the error will just be 0 and in this thesis our different space $\bar{V}_h$ involves piecewise polynomials of one degree higher than is used for V_h. That is if V_h involves piecewise polynomials of degree p then $\bar{V}_h$ involves piecewise polynomials of degree $p+1$. With this choice $\bar{V}_h$

is thus a larger space than V_h. Now in the representation in (4.2.7) we only know the quantities in the integrands when $s = 0$ and this leads to the following dual problem. Find $\underline{\psi}_h \in \bar{V}_h$ such that

$$A'(\underline{U}_h; \underline{\alpha}, \underline{\psi}_h) = J'(\underline{U}_h; \underline{\alpha}) \quad \forall \underline{\alpha} \in \bar{V}_h \tag{4.4.1}$$

which leads to the estimate

$$J(\underline{U}) - J(\underline{U}_h) \approx F(\underline{\psi}_h) - A(\underline{U}_h; \underline{\psi}_h). \tag{4.4.2}$$

The discussion about the existence and uniqueness of a solution that followed (4.2.7) similarly applies here and we comment later in this section about the accuracy of the estimate.

Although the problem just described is a linear problem it more computationally demanding than any step in a Newton iteration to get $\underline{U}_h$ as $\bar{V}_h$ is a larger function space than V_h. Now given the effort involved to get $\underline{\psi}_h$, we can also consider getting $\bar{\underline{U}}_h \in \bar{V}_h$ satisfying

$$A(\bar{\underline{U}}_h; \underline{\alpha}) = F(\underline{\alpha}) \quad \forall \underline{\alpha} \in \bar{V}_h. \tag{4.4.3}$$

In the quasi-static problems described later the approximation $\underline{U}_h \in V_h$ can be used to get the starting point in the Newton iteration used to get the coefficients of the approximation $\bar{\underline{U}}_h$. With $\bar{\underline{U}}_h$ usually being a better approximation to $\underline{U}$ than is $\underline{U}_h$ we can use it to create a dual problem which is closer to the exact dual problem as follows. By using the mid-point approximation rule, let

$$\underline{U}_h^m = \frac{1}{2}(\bar{\underline{U}}_h + \underline{U}_h) \tag{4.4.4}$$

and then let $\bar{\underline{\psi}}_h \in \bar{V}_h$ satisfy

$$A'\left(\underline{U}_h^m; \underline{\alpha}, \bar{\underline{\psi}}_h\right) = J'\left(\underline{U}_h^m; \underline{\alpha}\right) \quad \forall \underline{\alpha} \in \bar{V}_h. \tag{4.4.5}$$

The estimate of the QoI in this case is

$$J(\underline{U}) - J(\underline{U}_h) \approx F(\bar{\underline{\psi}}_h) - A(\underline{U}_h; \bar{\underline{\psi}}_h). \tag{4.4.6}$$

To explain why the estimate (4.4.6) is likely to be a good estimate when the mesh size h is sufficiently small and V_h and $\bar{V}_h$ involve polynomials of degree p and $p+1$ respectively can be done as follows given some assumptions about $A'(\cdot;\cdot,\cdot)$. We assume that $A'(\underline{U}_h + s\underline{e}_h; \cdot, \cdot)$ is coercive and bounded for all data in a region containing $\underline{U}_h$ and

$\underline{U}$, i.e. there exists constants $0 < c_0 \leq c_1$ such that

$$c_0 \left\|\underline{v}\right\|^2_{H^1(\Omega)} \leq \int_0^1 A'\left(\underline{U}_h + s\underline{e}_h; \underline{v}, \underline{v}\right)\, ds, \quad \forall \underline{v} \in V, \tag{4.4.7}$$

$$\left|\int_0^1 A'\left(\underline{U}_h + s\underline{e}_h; \underline{v}_1, \underline{v}_2\right)\, ds\right| \leq c_1 \left\|\underline{v}_1\right\|_{H^1(\Omega)} \left\|\underline{v}_2\right\|_{H^1(\Omega)} \quad \forall \underline{v}_1, \underline{v}_2 \in V, \tag{4.4.8}$$

where here $V = H^1_0(\Omega)$. As already stated the exact dual solution $\underline{\psi}$ satisfies (4.2.7) and as an intermediate problem to compare with, let $\tilde{\underline{\psi}}_h \in \bar{V}_h$ be such that

$$\int_0^1 A'\left(\underline{U}_h + s\underline{e}_h; \underline{\alpha}, \tilde{\underline{\psi}}_h\right)\, ds = \int_0^1 J'\left(\underline{U}_h + s\underline{e}_h; \underline{\alpha}\right)\, ds \quad \forall \underline{\alpha} \in \bar{V}_h. \tag{4.4.9}$$

When $\underline{\alpha} \in \bar{V}_h$ we can subtract (4.4.9) from (4.2.7) and use the linearity of the $A'(\cdot;\cdot,\cdot)$ terms after the semi-colon to write

$$\int_0^1 A'\left(\underline{U}_h + s\underline{e}_h; \underline{\alpha}, \underline{\psi} - \tilde{\underline{\psi}}_h\right)\, ds = 0 \quad \forall \underline{\alpha} \in \bar{V}_h. \tag{4.4.10}$$

We will use this Galerkin like orthogonality result in a moment. Now the coercive property with $\underline{\alpha} = \underline{\psi} - \tilde{\underline{\psi}}_h$ gives

$$c_0 \left\|\underline{\psi} - \tilde{\underline{\psi}}_h\right\|^2_{H^1(\Omega)} \leq \int_0^1 A'\left(\underline{U}_h + s\underline{e}_h; \underline{\psi} - \tilde{\underline{\psi}}_h, \underline{\psi} - \tilde{\underline{\psi}}_h\right)\, ds. \tag{4.4.11}$$

If we take the Galerkin orthogonality result (4.4.10) with $\underline{\alpha} = \underline{\psi}_h - \underline{\psi}_I$, where $\underline{\psi}_I \in \bar{V}_h$ is a suitably defined interpolant of $\underline{\psi} \in V$ from the finite element space $\bar{V}_h$, and add this to the right hand side of (4.4.11) then the inequality can be written as

$$\begin{aligned} c_0 \left\|\underline{\psi} - \tilde{\underline{\psi}}_h\right\|^2_{H^1(\Omega)} &\leq \int_0^1 A'\left(\underline{U}_h + s\underline{e}_h; \underline{\psi} - \underline{\psi}_I, \underline{\psi} - \tilde{\underline{\psi}}_h\right)\, ds & (4.4.12)\\ &\leq c_1 \left\|\underline{\psi} - \underline{\psi}_I\right\|_{H^1(\Omega)} \left\|\underline{\psi} - \tilde{\underline{\psi}}_h\right\|_{H^1(\Omega)}. & (4.4.13) \end{aligned}$$

Thus we get

$$\left\|\underline{\psi} - \tilde{\underline{\psi}}_h\right\|_{H^1(\Omega)} \leq \left(\frac{c_1}{c_0}\right) \left\|\underline{\psi} - \underline{\psi}_I\right\|_{H^1(\Omega)}, \tag{4.4.14}$$

which implies that

$$\left\|\underline{\psi} - \tilde{\underline{\psi}}_h\right\|_{H^1(\Omega)} = \mathcal{O}(h^{p+1}). \tag{4.4.15}$$

As the next intermediate problem we replace the exact error $\underline{e}_h$ by $\bar{\underline{e}}_h = \bar{\underline{U}}_h - \underline{U}_h$ and

let $\underline{\hat{\psi}}_h \in \bar{V}_h$ be the solution to

$$\int_0^1 A'\left(\underline{U}_h + s\underline{\bar{e}}_h; \underline{\alpha}, \underline{\hat{\psi}}_h\right) ds = \int_0^1 J'\left(\underline{U}_h + s\underline{\bar{e}}_h; \underline{\alpha}\right) ds \quad \forall \underline{\alpha} \in \bar{V}_h. \tag{4.4.16}$$

Both (4.4.9) and (4.4.16) involve the same test space $\bar{V}_h$ but they have different data and to compare them it helps to define $\delta_1(\underline{\alpha})$ and $\delta_2(\underline{\alpha})$ via the relations

$$\int_0^1 A'(\underline{U}_h + s\underline{e}_h; \underline{\alpha}, \tilde{\underline{\psi}}_h)\, ds = \int_0^1 A'(\underline{U}_h + s\underline{\bar{e}}_h; \underline{\alpha}, \tilde{\underline{\psi}}_h)\, ds + \delta_1(\underline{\alpha}), \tag{4.4.17}$$

$$\int_0^1 J'(\underline{U}_h + s\underline{e}_h; \underline{\alpha})\, ds = \int_0^1 J'(\underline{U}_h + s\underline{\bar{e}}_h; \underline{\alpha})\, ds + \delta_2(\underline{\alpha}). \tag{4.4.18}$$

By using the equation that each of $\tilde{\underline{\psi}}_h$ and $\hat{\underline{\psi}}_h$ satisfy we have

$$\begin{aligned}
\int_0^1 A'(\underline{U}_h + s\underline{\bar{e}}_h; \underline{\alpha}, \tilde{\underline{\psi}}_h)\, ds &= \int_0^1 A'(\underline{U}_h + s\underline{e}_h; \underline{\alpha}, \tilde{\underline{\psi}}_h)\, ds - \delta_1(\underline{\alpha}) && (4.4.19)\\
&= \int_0^1 J'(\underline{U}_h + s\underline{e}_h; \underline{\alpha})\, ds - \delta_1(\underline{\alpha}) && (4.4.20)\\
&= \int_0^1 J'(\underline{U}_h + s\underline{\bar{e}}_h; \underline{\alpha})\, ds + \delta_2(\underline{\alpha}) - \delta_1(\underline{\alpha}) && (4.4.21)\\
&= \int_0^1 A'(\underline{U}_h + s\underline{\bar{e}}_h; \underline{\alpha}, \hat{\underline{\psi}}_h)\, ds + \delta_2(\underline{\alpha}) - \delta_1(\underline{\alpha}). && (4.4.22)
\end{aligned}$$

By the linearity of $A(\cdot;\cdot,\cdot)$ in terms after the semi-colon we can collect the A' terms together and write

$$\int_0^1 A'(\underline{U}_h + s\underline{\bar{e}}_h; \underline{\alpha}, \tilde{\underline{\psi}}_h - \hat{\underline{\psi}}_h)\, ds = \delta_2(\underline{\alpha}) - \delta_1(\underline{\alpha}). \tag{4.4.23}$$

As the difference $\tilde{\underline{\psi}}_h - \hat{\underline{\psi}}_h \in \bar{V}_h$ the result holds with $\underline{\alpha} = \tilde{\underline{\psi}}_h - \hat{\underline{\psi}}_h \in \bar{V}_h$ and by using the coercivity property we have

$$\begin{aligned}
c_0 \left\| \tilde{\underline{\psi}}_h - \hat{\underline{\psi}}_h \right\|_{H^1}^2 &\leq \int_0^1 A'(\underline{U}_h + s\underline{\bar{e}}_h; \tilde{\underline{\psi}}_h - \hat{\underline{\psi}}_h, \tilde{\underline{\psi}}_h - \hat{\underline{\psi}}_h)\, ds\\
&= \delta_2(\tilde{\underline{\psi}}_h - \hat{\underline{\psi}}_h) - \delta_1(\tilde{\underline{\psi}}_h - \hat{\underline{\psi}}_h). && (4.4.24)
\end{aligned}$$

We separately bound the terms $\delta_1 = \delta_1(\tilde{\underline{\psi}}_h - \hat{\underline{\psi}}_h)$ and $\delta_2 = \delta_2(\tilde{\underline{\psi}}_h - \hat{\underline{\psi}}_h)$ after first giving expressions for each. By expressing the difference between two function values as by the

integral of the derivative we have

$$\begin{aligned}\delta_1 &= \int_0^1 A'(\underline{U}_h + s\underline{e}_h; \tilde{\underline{\psi}}_h - \hat{\underline{\psi}}_h, \tilde{\underline{\psi}}_h) - A'(\underline{U}_h + s\bar{\underline{e}}_h; \tilde{\underline{\psi}}_h - \hat{\underline{\psi}}_h, \tilde{\underline{\psi}}_h)\, ds \qquad (4.4.25)\\ &= \int_0^1 \int_0^1 A''(\underline{U}_h + s\underline{e}_h + ts(\underline{e}_h - \bar{\underline{e}}_h); s(\underline{e}_h - \bar{\underline{e}}_h), \tilde{\underline{\psi}}_h - \hat{\underline{\psi}}_h, \tilde{\underline{\psi}}_h)\, dtds \qquad (4.4.26)\end{aligned}$$

and

$$\begin{aligned}\delta_2 &= \int_0^1 J'(\underline{U}_h + s\underline{e}_h; \tilde{\underline{\psi}}_h - \hat{\underline{\psi}}_h) - J'(\underline{U}_h + s\bar{\underline{e}}_h; \tilde{\underline{\psi}}_h - \hat{\underline{\psi}}_h)\, ds \qquad (4.4.27)\\ &= \int_0^1 \int_0^1 J''(\underline{U}_h + s\underline{e}_h + ts(\underline{e}_h - \bar{\underline{e}}_h); s(\underline{e}_h - \bar{\underline{e}}_h), \tilde{\underline{\psi}}_h - \hat{\underline{\psi}}_h)\, dtds. \qquad (4.4.28)\end{aligned}$$

We assume that the A'' and J'' terms are bounded which specifically means that there are constants $c_2 > 0$ and $c_3 > 0$ such that for all $0 <\leq s, t \leq 1$ and for all $\underline{v}_1, \underline{v}_2, \underline{v}_3 \in H^1$

$$\begin{aligned}|A''(\underline{U}_h + s\underline{e}_h + t(\underline{e}_h - \bar{\underline{e}}_h); \underline{v}_1, \underline{v}_2, \underline{v}_3)| &\leq c_2 \|\underline{v}_1\|_{H^1} \|\underline{v}_2\|_{H^1} \|\underline{v}_3\|_{H^1}, \qquad (4.4.29)\\ |J''(\underline{U}_h + s\underline{e}_h + t(\underline{e}_h - \bar{\underline{e}}_h); \underline{v}_1, \underline{v}_2)| &\leq c_3 \|\underline{v}_1\|_{H^1} \|\underline{v}_2\|_{H^1}. \qquad (4.4.30)\end{aligned}$$

Then from how δ_1 and δ_2 are defined it follows that

$$\begin{aligned}|\delta_1(\tilde{\underline{\psi}}_h - \hat{\underline{\psi}}_h)| &\leq c_2 \|\underline{e}_h - \bar{\underline{e}}_h\|_{H^1} \left\|\tilde{\underline{\psi}}_h - \hat{\underline{\psi}}_h)\right\|_{H^1}, \qquad (4.4.31)\\ |\delta_2(\tilde{\underline{\psi}}_h - \hat{\underline{\psi}}_h)| &\leq c_3 \|\underline{e}_h - \bar{\underline{e}}_h\|_{H^1} \left\|\tilde{\underline{\psi}}_h - \hat{\underline{\psi}}_h)\right\|_{H^1}. \qquad (4.4.32)\end{aligned}$$

From these last two inequalities and (4.4.24) it follows that

$$\left\|\tilde{\underline{\psi}}_h - \hat{\underline{\psi}}_h\right\|_{H^1} \leq (\text{const}) \|\underline{e}_h - \bar{\underline{e}}_h\|_{H^1} = (\text{const}) \left\|\underline{U} - \bar{\underline{U}}_h\right\|_{H^1}. \qquad (4.4.33)$$

When the set-up of the problem defining $\underline{U}$ is sufficiently smooth such that a sequence of finite element approximations converge at their maximal rate it follows that when we have piecewise polynomials of degree $p+1$ for the space $\bar{V}_h$ we have

$$\left\|\tilde{\underline{\psi}}_h - \hat{\underline{\psi}}_h\right\|_{H^1} = \mathcal{O}(h^{p+1}). \qquad (4.4.34)$$

Finally we compare the function $\bar{\underline{\psi}}_h$ which satisfies (4.4.5) with the function $\hat{\underline{\psi}}_h$ which satisfies (4.4.16) which we have just considered. For the comparison we define $\delta_3(\underline{\alpha})$ and

$\delta_4(\underline{\alpha})$ via the relations

$$\int_0^1 A'(\underline{U}_h + s\bar{\underline{e}}_h; \underline{\alpha}, \hat{\underline{\psi}}_h)\, ds = A'(\underline{U}_h^m; \underline{\alpha}, \hat{\underline{\psi}}_h) + \delta_3(\underline{\alpha}), \tag{4.4.35}$$

$$\int_0^1 J'(\underline{U}_h + s\bar{\underline{e}}_h; \underline{\alpha})\, ds = J'(\underline{U}_h^m; \underline{\alpha}) + \delta_4(\underline{\alpha}). \tag{4.4.36}$$

By the equation satisfied by $\bar{\underline{\psi}}_h$ we can re-write (4.4.36) as

$$\int_0^1 J'(\underline{U}_h + s\bar{\underline{e}}_h; \underline{\alpha})\, ds = A'(\underline{U}_h^m; \underline{\alpha}, \bar{\underline{\psi}}_h) + \delta_4(\underline{\alpha}). \tag{4.4.37}$$

By the property of $\hat{\underline{\psi}}_h$ we can equate (4.4.35) and (4.4.37) to give

$$A'(\underline{U}_h^m; \underline{\alpha}, \hat{\underline{\psi}}_h) + \delta_3(\underline{\alpha}) = A'(\underline{U}_h^m; \underline{\alpha}, \bar{\underline{\psi}}_h) + \delta_4(\underline{\alpha}) \tag{4.4.38}$$

which, by the linearity of A' after the semi-colon, gives

$$A'(\underline{U}_h^m; \underline{\alpha}, \hat{\underline{\psi}}_h - \bar{\underline{\psi}}_h) = \delta_4(\underline{\alpha}) - \delta_3(\underline{\alpha}). \tag{4.4.39}$$

Now if we take $\underline{\alpha} = \hat{\underline{\psi}}_h - \bar{\underline{\psi}}_h \in \bar{V}_h$ and use coercivity we have

$$c_0 \left\| \hat{\underline{\psi}}_h - \bar{\underline{\psi}}_h \right\|_{H^1}^2 \leq A'(\underline{U}_h^m; \hat{\underline{\psi}}_h - \bar{\underline{\psi}}_h, \hat{\underline{\psi}}_h - \bar{\underline{\psi}}_h) = \delta_4(\hat{\underline{\psi}}_h - \bar{\underline{\psi}}_h) - \delta_3(\hat{\underline{\psi}}_h - \bar{\underline{\psi}}_h). \tag{4.4.40}$$

We separately bound the δ_3 and δ_4 terms next which are both concerned with the error in the mid-point rule approximation of an integral.

Now for a 2-times continuously differentiable function $\phi : [0,1] \to \mathbb{R}$ Taylor's series with remainder gives

$$\phi(s) = \phi(1/2) + \phi'(1/2)(s - 1/2) + \frac{1}{2}\phi''(\xi(s))(s - 1/2)^2$$

for some $0 < \xi(s) < 1$. It then follows that

$$\begin{aligned} \int_0^1 (\phi(s) - \phi(1/2))\, ds &= \frac{1}{2}\int_0^1 \phi''(\xi(s))(s - 1/2)^2\, ds \\ &= \frac{1}{2}\phi''(\eta)\int_0^1 (s - 1/2)^2\, ds = \frac{1}{24}\phi''(\eta) \end{aligned} \tag{4.4.41}$$

for some $0 < \eta < 1$. The main point here is that the mid-point rule is exact for degree 0 and 1 polynomials and the error depends on the second derivative. In the case of

bounding $\delta_3(\underline{\hat{\psi}}_h - \underline{\bar{\psi}}_h)$ the function $\phi(s)$ is

$$\phi(s) = A'(\underline{U}_h + s\underline{\bar{e}}_h; \underline{\hat{\psi}}_h - \underline{\bar{\psi}}_h, \underline{\hat{\psi}}_h)$$

and, in terms of higher Gâteaux derivatives,

$$\phi''(s) = A'''(\underline{U}_h + s\underline{\bar{e}}_h; \underline{\bar{e}}_h, \underline{\bar{e}}_h, \underline{\hat{\psi}}_h - \underline{\bar{\psi}}_h, \underline{\hat{\psi}}_h). \tag{4.4.42}$$

Similarly in the case of bounding $\delta_4(\underline{\hat{\psi}}_h - \underline{\bar{\psi}}_h)$ the function $\phi(s)$ is

$$\phi(s) = J'(\underline{U}_h + s\underline{\bar{e}}_h; \underline{\hat{\psi}}_h - \underline{\bar{\psi}}_h)$$

and in this case

$$\phi''(s) = J'''(\underline{U}_h + s\underline{\bar{e}}_h; \underline{\bar{e}}_h, \underline{\bar{e}}_h, \underline{\hat{\psi}}_h - \underline{\bar{\psi}}_h). \tag{4.4.43}$$

We assume that A''' and J''' are bounded which specifically means that there are constants $c_4 > 0$ and $c_5 > 0$ such that for all $0 \le s \le 1$ and for all $\underline{v}_1, \underline{v}_2, \underline{v}_3, \underline{v}_4 \in H^1$

$$|A'''(\underline{U}_h + s\underline{e}_h; \underline{v}_1, \underline{v}_2, \underline{v}_3, \underline{v}_4)| \le c_4 \|\underline{v}_1\|_{H^1} \|\underline{v}_2\|_{H^1} \|\underline{v}_3\|_{H^1} \|\underline{v}_4\|_{H^1}, \tag{4.4.44}$$

$$|J'''(\underline{U}_h + s\underline{e}_h + t(\underline{e}_h - \underline{\bar{e}}_h); \underline{v}_1, \underline{v}_2, \underline{v}_3)| \le c_5 \|\underline{v}_1\|_{H^1} \|\underline{v}_2\|_{H^1} \|\underline{v}_3\|_{H^1}. \tag{4.4.45}$$

Using this assumption with the definitions of δ_3 and δ_4 and the result about the error in the mid-point rule gives the inequalities

$$|\delta_3(\underline{\hat{\psi}}_h - \underline{\bar{\psi}}_h)| \le c_4 \|\underline{\bar{e}}_h\|_{H^1}^2 \left\|\underline{\hat{\psi}}_h - \underline{\bar{\psi}}_h\right\|_{H^1}, \tag{4.4.46}$$

$$|\delta_4(\underline{\hat{\psi}}_h - \underline{\bar{\psi}}_h)| \le c_5 \|\underline{\bar{e}}_h\|_{H^1}^2 \left\|\underline{\hat{\psi}}_h - \underline{\bar{\psi}}_h\right\|_{H^1}. \tag{4.4.47}$$

Using these last 2 inequalities in (4.4.40) gives

$$\left\|\underline{\hat{\psi}}_h - \underline{\bar{\psi}}_h\right\|_{H^1} \le (\text{const}) \|\underline{\bar{e}}\|_{H^1}^2 \tag{4.4.48}$$

and hence

$$\left\|\underline{\hat{\psi}}_h - \underline{\bar{\psi}}_h\right\|_{H^1} = \mathcal{O}(h^{2p}). \tag{4.4.49}$$

As

$$\underline{\psi} - \underline{\bar{\psi}}_h = (\underline{\psi} - \underline{\tilde{\psi}}_h) + (\underline{\tilde{\psi}}_h - \underline{\hat{\psi}}_h) + (\underline{\hat{\psi}}_h - \underline{\bar{\psi}}_h)$$

the triangle inequality and $p + 1 \le 2p$ for $p \ge 1$ gives

$$\left\|\underline{\psi} - \underline{\bar{\psi}}_h\right\|_{H^1} \le \left\|\underline{\psi} - \underline{\tilde{\psi}}_h\right\|_{H^1} + \left\|\underline{\tilde{\psi}}_h - \underline{\hat{\psi}}_h\right\|_{H^1} + \left\|\underline{\hat{\psi}}_h - \underline{\bar{\psi}}_h\right\|_{H^1} = \mathcal{O}(h^{p+1}). \tag{4.4.50}$$

The other dual problem given in (4.4.1) computes instead $\underline{\psi}_h$ and to consider how well this approximates $\underline{\psi}$ can be done by replacing the mid-point rule step above with the left hand approximation rule to the integral the accuracy is less. Firstly, when $s > 0$,

$$\phi(s) = \phi(0) + \phi'(\xi(s))s$$

for some $0 < \xi(s) < s < 1$ and

$$\int_0^1 \phi(s) - \phi(0)\, ds = \int_0^1 \phi'(\xi(s))s\, ds = \phi'(\eta) \int_0^1 s\, ds = \frac{1}{2}\phi'(\eta) \tag{4.4.51}$$

for some $0 < \eta < 1$. The rule is exact for degree 0 polynomials but not for degree 1 polynomials. Instead of getting the terms (4.4.42) and (4.4.43) we have to consider

$$A''(\underline{U}_h + s\underline{\bar{e}}_h; \underline{\bar{e}}_h, \underline{\hat{\psi}}_h - \underline{\bar{\psi}}_h, \underline{\hat{\psi}}_h) \quad \text{and} \quad J''(\underline{U}_h + s\underline{\bar{e}}_h; \underline{\bar{e}}_h, \underline{\hat{\psi}}_h - \underline{\bar{\psi}}_h).$$

and for the corresponding terms δ_3 and δ_4 we get for some constants $c_6 > 0$ and $c_7 > 0$ that

$$|\delta_3(\underline{\hat{\psi}}_h - \underline{\bar{\psi}}_h)| \leq c_6 \|\underline{\bar{e}}_h\|_{H^1} \left\|\underline{\hat{\psi}}_h - \underline{\bar{\psi}}_h\right\|_{H^1}, \tag{4.4.52}$$

$$|\delta_4(\underline{\hat{\psi}}_h - \underline{\bar{\psi}}_h)| \leq c_7 \|\underline{\bar{e}}_h\|_{H^1} \left\|\underline{\hat{\psi}}_h - \underline{\bar{\psi}}_h\right\|_{H^1}. \tag{4.4.53}$$

Thus instead of (4.4.48) we get

$$\left\|\underline{\hat{\psi}}_h - \underline{\psi}_h\right\|_{H^1(\Omega)} \leq (\text{const}) \|\underline{\bar{e}}_h\|_{H^1(\Omega)} = O(h^p), \tag{4.4.54}$$

so that overall we have

$$\left\|\underline{\psi} - \underline{\psi}_h\right\|_{H^1(\Omega)} = \mathcal{O}(h^p). \tag{4.4.55}$$

We consider now what the results (4.4.50) and (4.4.55) mean for our estimates. In the case of $\underline{\psi}_h$ we have

$$\begin{aligned}
& J(\underline{U}) - J(\underline{U}_h) - \Big(F(\underline{\psi}_h) - A(\underline{U}_h; \underline{\psi}_h)\Big) \\
= \; & \big(F(\underline{\psi}) - A(\underline{U}_h; \underline{\psi})\big) - \Big(F(\underline{\psi}_h) - A(\underline{U}_h; \underline{\psi}_h)\Big) \\
= \; & F(\underline{\psi} - \underline{\psi}_h) - A(\underline{U}_h; \underline{\psi} - \underline{\psi}_h) \\
= \; & A(\underline{U}; \underline{\psi} - \underline{\psi}_h) - A(\underline{U}_h; \underline{\psi} - \underline{\psi}_h) \\
= \; & \int_0^1 A'(\underline{U}_h + s\underline{e}_h; \underline{e}_h, \underline{\psi} - \underline{\psi}_h)\, \mathrm{d}s.
\end{aligned} \tag{4.4.56}$$

When $A'(\underline{U}_h + s\underline{e}_h; \cdot, \cdot)$ is bounded for all $s \in [0,1]$, the expression involves a product of terms in the components of $\underline{e}_h$ and $\underline{\psi} - \underline{\psi}_h$ and this leads to

$$J(\underline{U}) - J(\underline{U}_h) - \Big(F(\underline{\psi}_h) - A(\underline{U}_h; \underline{\psi}_h)\Big) = \mathcal{O}(h^{2p}), \tag{4.4.57}$$

since $\|\underline{e}_h\|_{H^1}$ and $\left\|\underline{\psi} - \underline{\psi}_h\right\|_{H^1}$ are both $O(h^p)$. The estimate $F(\underline{\psi}_h) - A(\underline{U}_h; \underline{\psi}_h)$ hence tends to 0 at the same rate as $J(\underline{U}) - J(\underline{U}_h)$. Similar reasoning in the case of $\bar{\underline{\psi}}_h$ gives

$$J(\underline{U}) - J(\underline{U}_h) - \Big(F(\bar{\underline{\psi}}_h) - A(\underline{U}_h; \bar{\underline{\psi}}_h)\Big) = \int_0^1 A'(\underline{U}_h + s\underline{e}_h; \underline{e}_h, \underline{\psi} - \bar{\underline{\psi}}_h)\, \mathrm{d}s. \tag{4.4.58}$$

In this case the expression involves a product of terms in the components of $\underline{e}_h$ and $\underline{\psi} - \bar{\underline{\psi}}_h$ and this leads to

$$J(\underline{U}) - J(\underline{U}_h) - \Big(F(\bar{\underline{\psi}}_h) - A(\underline{U}_h; \bar{\underline{\psi}}_h)\Big) = \mathcal{O}(h^{2p+1}). \tag{4.4.59}$$

Thus we have asymptotic exactness when we go the extra computational expense of obtaining $\bar{\underline{\psi}}_h$.

4.5 Comments about a Taylor's series representation of $J(\underline{U}) - J(\underline{U}_h)$

For the computational problems considered later we have already given the results that we need in an abstract setting about the dual problems that are going to be considered. As background it is useful to present one further result which is often given as a way of representing $J(\underline{U}) - J(\underline{U}_h)$ by using a Taylor series type expression with a remainder, see e.g. Oden and Prudhomme in [17]. The result involves yet another dual problem which uses the exact solution $\underline{U}$ as data. The result can be described as follows.

Let $\underline{\psi}_U \in V$ satisfy

$$A'(\underline{U}; \underline{\alpha}, \underline{\psi}_U) = J'(\underline{U}; \underline{\alpha}) \quad \forall \underline{\alpha} \in V \tag{4.5.1}$$

and to repeat what has been given before let $\underline{\psi}_h \in \bar{V}_h$ satisfy

$$A'(\underline{U}_h; \underline{\alpha}, \underline{\psi}_h) = J'(\underline{U}_h; \underline{\alpha}) \quad \forall \underline{\alpha} \in \bar{V}_h. \tag{4.5.2}$$

Also, as before, let

$$\underline{e}_h = \underline{U} - \underline{U}_h \quad \text{and now define} \quad \underline{e}_\psi = \underline{\psi}_U - \underline{\psi}_h. \tag{4.5.3}$$

If we let

$$g_0(s) = J(\underline{U}_h + s\underline{e}_h) + F(\underline{\psi}_h + s\underline{e}_\psi) - A(\underline{U}_h + s\underline{e}_h; \underline{\psi}_h + s\underline{e}_\psi) \tag{4.5.4}$$

then the derivative of this with respect to s is

$$g_0'(s) = J'(\underline{U}_h + s\underline{e}_h; \underline{e}_h) + F(\underline{e}_\psi) - A'(\underline{U}_h + s\underline{e}_h; \underline{e}_h, \underline{\psi}_h + s\underline{e}_\psi) - A(\underline{U}_h + s\underline{e}_h; \underline{e}_\psi). \tag{4.5.5}$$

Then we get the following

$$\begin{aligned} g_0(1) &= J(\underline{U}) \quad \text{and} \quad g_0(0) = J(\underline{U}_h) + F(\underline{\psi}_h) - A(\underline{U}_h; \underline{\psi}_h) && (4.5.6)\\ g_0'(1) &= 0 \quad \text{and} \quad g_0'(0) = J'(\underline{U}_h; \underline{e}_h) + F(\underline{e}_\psi) - A'(\underline{U}_h; \underline{e}_h, \underline{\psi}_h) - A(\underline{U}_h; \underline{e}_\psi). && (4.5.7) \end{aligned}$$

Using (4.5.6) we get

$$J(\underline{U}) - J(\underline{U}_h) = F(\underline{\psi}_h) - A(\underline{U}_h; \underline{\psi}_h) + g_0(1) - g_0(0). \tag{4.5.8}$$

This brings in the estimate in the first scheme described and the difference between the true value and the estimate can be written as

$$g_0(1) - g_0(0) = \int_0^1 g_0'(s)\, ds = \frac{1}{2}(g_0'(0) + g_0'(1)) + \frac{1}{2}\int_0^1 s(s-1) g_0'''(s)\, ds \tag{4.5.9}$$

by using the trapezoidal rule with an integral form of the remainder. The term involving $g_0'''(s)$ involves the product of 3 terms. Now, by using (4.5.7) and with a little manipulation, it can be shown that

$$\frac{1}{2}(g_0'(0) + g_0'(1)) = \big(F(\underline{e}_\psi) - A(\underline{U}_h; \underline{e}_\psi)\big) + \chi \tag{4.5.10}$$

where

$$\chi = \frac{1}{2}\int_0^1 J''(\underline{U}_h + s\underline{e}_h; \underline{e}_h, \underline{e}_h) - A''(\underline{U}_h + s\underline{e}_h; \underline{e}_h, \underline{e}_h, \underline{\psi}_h + s\underline{e}_\psi) - A'(\underline{U}_h + s\underline{e}_h; \underline{e}_h, \underline{e}_\psi)\, ds. \tag{4.5.11}$$

By using the equation (4.5.1) the first term

$$\begin{aligned} F(\underline{e}_\psi) - A(\underline{U}_h; \underline{e}_\psi) &= A(\underline{U}; \underline{e}_\psi) - A(\underline{U}_h; \underline{e}_\psi) && (4.5.12)\\ &= \int_0^1 A'(\underline{U}_h + s(\underline{U} - \underline{U}_h); \underline{U} - \underline{U}_h; \underline{e}_\psi)\, ds. && (4.5.13) \end{aligned}$$

Assuming that $A'(\cdot;\cdot,\cdot)$ is bounded implies that

$$|F(\underline{e}_\psi) - A(\underline{U}_h; \underline{e}_\psi)| \le (\text{const})\, \|\underline{U} - \underline{U}_h\|_{H^1} \left\|\underline{e}_\psi\right\|_{H^1}. \tag{4.5.14}$$

When we use degree p polynomials we have that $\|\underline{U} - \underline{U}_h\|_{H^1}$ is $\mathcal{O}(h^p)$. The reasoning of section 4.4 shows that when we change the data in the dual problem by $\mathcal{O}(h^p)$ the solution changes by this order in the H^1 norm and thus $\left\|\underline{e}_\psi\right\|_{H^1(\Omega)}$ is also $\mathcal{O}(h^p)$. Thus the expression in (4.5.10) decays like $\mathcal{O}(h^{2p})$ which is what has already been deduced.

4.6 A dual problem for the non-axisymmetric inflation problem

In this section we give more computational details for the description of the dual problem for the non-axisymmetric inflation problem for the quasi-static case. We describe how we solve the problem using the element-by-element way, similarly with the weak problem, and how we apply the Gâteax derivatives of the weak form $A(\cdot;\cdot)$ and the functional $J(\cdot)$ for this case.

In the earlier parts of this chapter we outlined the set-up in an abstract way in creating a dual problem to be used to assess the accuracy in an estimate $J(\underline{u}_h)$ of a QoI $J(\underline{u})$ where J denotes the QoI functional. At that stage it was mentioned that the details in any given case can be highly problem dependent and we consider here these details in the case of the quasi-static inflation of a non-axisymmetric membrane which is the topic of this chapter.

In (4.4.1) and (4.4.5) we gave the dual problems to be considered in a computation and for the description here we let $\underline{\tilde{u}}_h$ be where we evaluate the expressions. That is, $\underline{\tilde{u}}_h$ is $\underline{u}_h$ or it is the average of $\underline{u}_h$ and a better approximation and its role here is as the data for the dual problem. The complete expressions involve an integral over the domain Ω and we get contributions to this from each element Ω_r which we consider in a moment. When these element quantities have been determined and assembled the dual solution $\underline{\psi}$ in a space $\bar{V}_h$ is such that

$$A'(\underline{\tilde{u}}_h; \underline{\alpha}, \underline{\psi}) = J'(\underline{\tilde{u}}_h; \underline{\alpha}), \quad \forall \underline{\alpha} \in \bar{V}_h. \tag{4.6.1}$$

With piecewise linears being used to get $\underline{u}_h$ we use 6-noded quadratics on the same triangular mesh of Ω which defines the space $\bar{V}_h$. We set-up the finite element calculation for (4.6.1) in an element-by-element way and we consider next the element contribution to the left hand side in (4.6.1). Later we indicate the expressions for $J'(\cdot;\cdot)$ for two QoI consider in this thesis.

4.6.1 The element matrix for the dual problem

On an element Ω_r the dual solution $\underline{\psi}$ can be described in the form

$$\underline{\psi}(x_1, x_2) = \sum_{i=1}^{m} c_i \underline{\hat{\phi}}_i(x_1, x_2) \tag{4.6.2}$$

with $\underline{\hat{\phi}}_1, \ldots, \underline{\hat{\phi}}_m$ being the basis functions on an element and for the 6-noded quadratics $m = 18$. Corresponding to a_1 and a_2 given in (3.6.12) and (3.6.13) we define

$$A'(\underline{\tilde{u}}_h; \underline{\hat{\phi}}_j, \underline{\hat{\phi}}_i)_r = a_1'(\underline{\tilde{u}}_h; \underline{\hat{\phi}}_j, \underline{\hat{\phi}}_i)_{\Omega_r} - P a_2'(\underline{\tilde{u}}_h; \underline{\hat{\phi}}_j, \underline{\hat{\phi}}_i)_{\Omega_r}, \quad 1 \le i, j \le m, \tag{4.6.3}$$

where $a_1'(\underline{\tilde{u}}_h; \underline{\hat{\phi}}_j, \underline{\hat{\phi}}_i)_{\Omega_r}$ and $a_2'(\underline{\tilde{u}}_h; \underline{\hat{\phi}}_j, \underline{\hat{\phi}}_i)_{\Omega_r}$ are respectively the Gâteaux derivatives in the direction of $\underline{\hat{\phi}}_j$ of

$$a_1(\underline{u}_h; \underline{\hat{\phi}}_i)_{\Omega_r} = h_0 \iint_{\Omega_r} \frac{\partial W}{\partial \mathbf{F}} : \nabla \underline{\hat{\phi}}_i \, dx_1 dx_2, \tag{4.6.4}$$

$$\begin{aligned} a_2(\underline{u}_h; \underline{\hat{\phi}}_i)_{\Omega_r} = {} & \frac{1}{3} \iint_{\Omega_r} \underline{\hat{\phi}}_i \cdot \left(\frac{\partial \underline{w}}{\partial x_1} \times \frac{\partial \underline{w}}{\partial x_2} \right) dx_1 dx_2 + \frac{1}{3} \iint_{\Omega_r} \underline{w} \cdot \left(\frac{\partial \underline{\hat{\phi}}_i}{\partial x_1} \times \frac{\partial \underline{w}}{\partial x_2} \right) dx_1 dx_2 \\ & + \frac{1}{3} \iint_{\Omega_r} \underline{w} \cdot \left(\frac{\partial \underline{w}}{\partial x_1} \times \frac{\partial \underline{\hat{\phi}}_i}{\partial x_2} \right) dx_1 dx_2 \end{aligned} \tag{4.6.5}$$

Apart from how functions are labelled, we have effectively already given what is needed to evaluate the Gâteaux derivatives from the description given to determine the element Jacobian contributions in the Newton iteration. In (3.6.21) and (3.6.22) the expressions correspond to changes to the finite element approximation in the direction of the basis function $\underline{\phi}_j$ used at that stage. In the dual problem considered here the direction now corresponds to the basis functions $\underline{\hat{\phi}}_j$ in the larger finite element space $\bar{V}_h$. These expressions in the dual problem case are as follows.

$$a_1'(\underline{\tilde{u}}_h; \underline{\hat{\phi}}_j, \underline{\hat{\phi}}_i) = h_0 \iint_{\Omega_r} \sum_{r=1}^{6} \sum_{s=1}^{6} \frac{\partial^2 W}{\partial F_s \partial F_r} \left(\bar{\nabla} \underline{\hat{\phi}}_i \right)_r \left(\bar{\nabla} \underline{\hat{\phi}}_j \right)_s dx_1 dx_2 \tag{4.6.6}$$

and

$$\begin{aligned} a_2'(\tilde{\underline{u}}_h;\hat{\underline{\phi}}_j,\hat{\underline{\phi}}_i) \;=\; & \iint_{\Omega_r} \hat{\underline{\phi}}_i \cdot \left(\frac{\partial \hat{\underline{\phi}}_j}{\partial x_1} \times \frac{\partial \underline{w}}{\partial x_2} + \frac{\partial \underline{w}}{\partial x_1} \times \frac{\partial \hat{\underline{\phi}}_j}{\partial x_2} \right) \\ & + \underline{w} \cdot \left(\frac{\partial \hat{\underline{\phi}}_i}{\partial x_1} \times \frac{\partial \hat{\underline{\phi}}_j}{\partial x_2} + \frac{\partial \hat{\underline{\phi}}_j}{\partial x_1} \times \frac{\partial \hat{\underline{\phi}}_i}{\partial x_2} \right) \\ & + \hat{\underline{\phi}}_j \cdot \left(\frac{\partial \hat{\underline{\phi}}_i}{\partial x_1} \times \frac{\partial \underline{w}}{\partial x_2} + \frac{\partial \underline{w}}{\partial x_1} \times \frac{\partial \hat{\underline{\phi}}_i}{\partial x_2} \right) dx_1 dx_2 \end{aligned} \tag{4.6.7}$$

4.6.2 Examples of J and the J' expression

The right hand side in (4.6.1) involves the Gâteaux derivative of the QoI functional J being considered and we can handle expressions of the form

$$J(\underline{u}) = \iint_{\Omega^*} \left\{ \text{expressions in } \underline{u},\, \frac{\partial \underline{u}}{\partial x_1},\, \frac{\partial \underline{u}}{\partial x_2} \right\} dx_1 dx_2 \tag{4.6.8}$$

where $\Omega^* \subset \Omega$. Two cases of this type are the following.

1. $$J_1(\underline{u}) = \frac{1}{\text{area}(\Omega^*)} \iint_{\Omega^*} \lambda \, dx_1 dx_2 \tag{4.6.9}$$

   where $\Omega^* \subset \Omega$. The above expression gives **the average thickness stretch ratio over the domain** Ω^*.

2. $$J_2(\underline{u}) = h_0 \iint_{\Omega} W \, dx_1 dx_2 - \frac{P}{3} \iint_{\Omega} \underline{w} \cdot \left(\frac{\partial \underline{w}}{\partial x_1} \times \frac{\partial \underline{w}}{\partial x_2} \right) dx_1 dx_2. \tag{4.6.10}$$

   The above expression gives **the potential energy of the deformed membrane**.

We complete this section and this chapter by deriving the expressions for J_1' and J_2'.

The Gâteaux derivative of J_1

As has already shown in section 3.4.1, the incompressibility assumption can be written as

$$\det(\mathbf{C}_{3D}) = \lambda^2 (c_{11}c_{22} - c_{21}^2) = 1 \quad \text{and thus} \quad \lambda = (c_{11}c_{22} - c_{21}^2)^{-1/2}. \tag{4.6.11}$$

Recall that the first Gâteaux derivative of λ in the direction of $\underline{\alpha}$ is defined by

$$\lambda'(\underline{u};\underline{\alpha}) = \frac{\mathrm{d}}{\mathrm{d}s}\lambda(\underline{u}+s\underline{\alpha})\bigg|_{s=0}. \tag{4.6.12}$$

Each of c_{11}, c_{22} and c_{12} depend on the partial derivatives of $\underline{u}$. Thus by combining (4.6.11) and (4.6.12) we get

$$\begin{aligned}\lambda'(\underline{u};\underline{\alpha}) &= -\frac{1}{2}\left(c_{11}c_{22}-c_{21}^2\right)^{-3/2}\left(c_{11}c'_{22}(\underline{u};\underline{\alpha})+c'_{11}(\underline{u};\underline{\alpha})c_{22}-2c_{21}c'_{21}(\underline{u};\underline{\alpha})\right)\\ &= -\frac{\lambda^3}{2}\left(c_{11}c'_{22}(\underline{u};\underline{\alpha})+c'_{11}(\underline{u};\underline{\alpha})c_{22}-2c_{12}c'_{12}(\underline{u};\underline{\alpha})\right).\end{aligned}$$

Let $\underline{f}_1$ and $\underline{f}_2$ denote respectively columns 1 and 2 of $\mathbf{F}$ and recall that $\mathbf{F}$ is given by

$$\mathbf{F} = \begin{pmatrix}1&0\\0&1\\0&0\end{pmatrix} + \nabla\underline{u} = \begin{pmatrix}1&0\\0&1\\0&0\end{pmatrix} + \left(\frac{\partial\underline{u}}{\partial x_1},\frac{\partial\underline{u}}{\partial x_2}\right).$$

Now as

$$c_{11} = \underline{f}_1^T\underline{f}_1, \quad c_{22} = \underline{f}_2^T\underline{f}_2 \quad \text{and} \quad c_{12} = \underline{f}_1^T\underline{f}_2$$

it follows that

$$c'_{11}(\underline{u};\underline{\alpha}) = 2\underline{f}_1^T\frac{\partial\underline{\alpha}}{\partial x_1}, \quad c'_{22}(\underline{u};\underline{\alpha}) = 2\underline{f}_2^T\frac{\partial\underline{\alpha}}{\partial x_2}, \quad \text{and} \quad c'_{12}(\underline{u};\underline{\alpha}) = \underline{f}_1^T\frac{\partial\underline{\alpha}}{\partial x_2}+\underline{f}_2^T\frac{\partial\underline{\alpha}}{\partial x_1}. \tag{4.6.13}$$

Putting things together gives

$$\lambda'(\underline{u};\underline{\alpha}) = -\lambda^3\left((c_{22}\underline{f}_1^T - c_{12}\underline{f}_2^T)\frac{\partial\underline{\alpha}}{\partial x_1} + (c_{11}\underline{f}_2^T - c_{12}\underline{f}_1^T)\frac{\partial\underline{\alpha}}{\partial x_2}\right). \tag{4.6.14}$$

The Gâteaux derivative of J_2

To get the Gâteaux derivative of J_2 let

$$\Psi = h_0W - \frac{P}{3}\left(\underline{w}\cdot\left(\frac{\partial\underline{w}}{\partial x_1}\times\frac{\partial\underline{w}}{\partial x_2}\right)\right), \quad \text{where } \underline{w} = \begin{pmatrix}x_1\\x_2\\0\end{pmatrix} + \begin{pmatrix}u_1\\u_2\\u_3\end{pmatrix}. \tag{4.6.15}$$

When we consider W in the form $W = W(\mathbf{F})$ it follows that

$$\frac{\mathrm{d}}{\mathrm{d}s}W(\mathbf{F}(\underline{u}+s\underline{\alpha}))\bigg|_{s=0} = \frac{\partial W}{\partial\mathbf{F}} : \nabla\underline{\alpha}. \tag{4.6.16}$$

Thus by the product rule

$$\Psi'(\underline{u};\underline{\alpha}) = h_0 \frac{\partial W}{\partial \mathbf{F}} : \nabla \underline{\alpha} \\ -\frac{P}{3}\left(\underline{\alpha}\cdot\left(\frac{\partial \underline{w}}{\partial x_1}\times\frac{\partial \underline{w}}{\partial x_2}\right)+\underline{w}\cdot\left(\frac{\partial \underline{\alpha}}{\partial x_1}\times\frac{\partial \underline{w}}{\partial x_2}\right)+\underline{w}\cdot\left(\frac{\partial \underline{w}}{\partial x_1}\times\frac{\partial \underline{\alpha}}{\partial x_2}\right)\right). \quad (4.6.17)$$

By comparing with (3.4.15)–(3.4.18) this shows that

$$J_2'(\underline{u};\underline{\alpha}) = A(\underline{u};\underline{\alpha}), \quad (4.6.18)$$

the expression in the weak form which defines $\underline{u}$.

5. RESULTS WITH THE MEMBRANE MODEL FOR THE QUASI-STATIC CASE

5.1 Introduction and the problems

In this chapter we give some numerical examples to test the theory presented in the previous chapter concerning when the method involving solving a dual problem enables us to determine the error in a quantity of interest. We consider the two functionals given in section 4.6.2 involving the average thickness over part of the domain and involving the potential energy of the deformation. J_1 denotes the average thickness and J_2 denotes the potential energy. We consider two different domains Ω for the region of the undeformed mid-surface and these are a square and a L-shape. In all the examples the Mooney-Rivlin strain energy function, which we express in in the Ogden form (3.5.22),

$$\begin{aligned} W &= \left(\lambda_1^2+\lambda_2^2+\lambda_3^2-3\right)+\left(\frac{-0.1}{-2}\right)\left(\lambda_1^{-2}+\lambda_2^{-2}+\lambda_3^{-2}-3\right) \\ &= \left(\lambda_1^2+\lambda_2^2+\lambda_3^2-3\right)+0.05\left(\lambda_1^{-2}+\lambda_2^{-2}+\lambda_3^{-2}-3\right) \end{aligned} \tag{5.1.1}$$

is used and we start with a prestretch such that the membrane deformation gradient is

$$\mathbf{F}=\begin{pmatrix} 1.2 & 0 \\ 0 & 1.2 \\ 0 & 0 \end{pmatrix}. \tag{5.1.2}$$

To complete the description for each case of Ω we need to indicate the pressure P involved and we select this to get a moderate amount of stretching based on the solutions obtained with the "first mesh" used in the computation. More specifically we do the following.

When Ω is a square

When Ω is a square a symmetrical mesh of 8 triangles is given in figure 5.1(a) This is uniformly refined 2 times to give the mesh of $8\times4\times4 = 128$ elements shown in figure 5.1(b) and this is the first mesh in the computations. We then solve at pressures $P = 0.05$, 0.1, $0.15\ldots$, until we get a solution with $\max\{u_3(x_1, x_2) : \ (x_1, x_2) \in \Omega\} > 1$. This first occurs with $P = 1.8$ and we use this pressure with the finer meshes. The deformed membrane at this stage is shown in figure 5.3(a).

When Ω is a L-shape

When Ω is a L-shape we can describe the domain with 6 triangles as shown in figure 5.2(a). This mesh is then uniformly refined 3 times to give the mesh of $6 \times 4 \times 4 \times 4 = 384$ elements shown in figure 5.2(b) and this the "first mesh" for this geometry in the computations. We then solve at pressures $P = 0.05$, 0.1, $0.15\ldots$, until we get a solution with $\max\{u_3(x_1, x_2) : \ (x_1, x_2) \in \Omega\} > 1$ and this first occurs with $P = 3.2$ and we use this pressure with all the finer meshes. The deformed membrane at this stage is shown in figure 5.3(b).

The regions Ω^* when Ω is a square

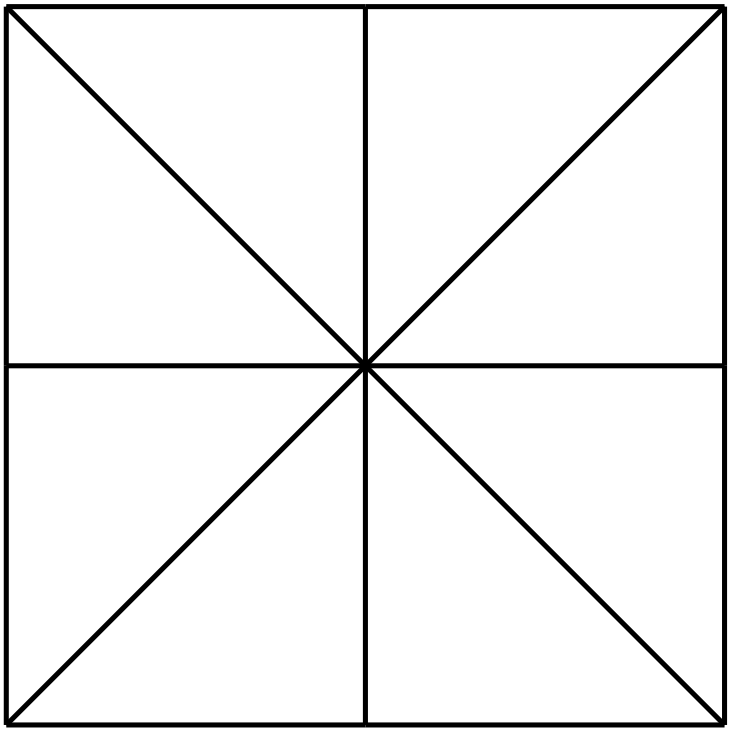
With Ω just defined it is convenient at this point to introduce the parts of Ω which are used in the next section when we define the quantities of interest that are considered and in particular Ω^* is the part of Ω where we consider the average thickness over this part as is given in (4.6.8). When we consider a part Ω^* of the square $\Omega = \{(x_1, x_2) : \ -1 \le x_1, x_2 \le 1\}$ we take the following two cases

$$\Omega_1^* \ = \ \{(x_1, x_2) : \ -0.25 \le x_1, x_2 \le 0.25\} \tag{5.1.3}$$

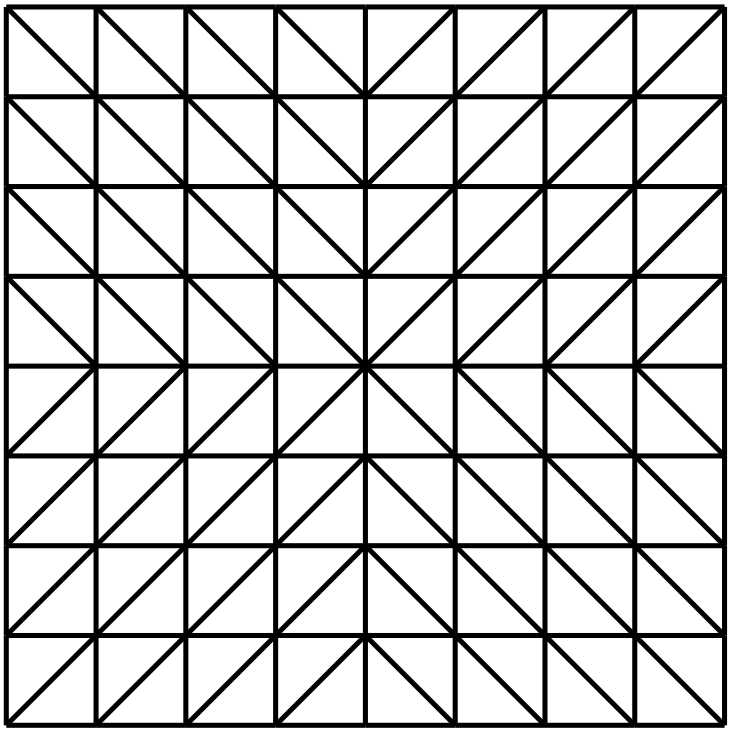
and

$$\tag{5.1.4}$$

$$\Omega_2^* \ = \ \{(x_1, x_2) : \ |x_1| \ge 0.75, \ |x_2| \ge 0.75\}\,. \tag{5.1.5}$$

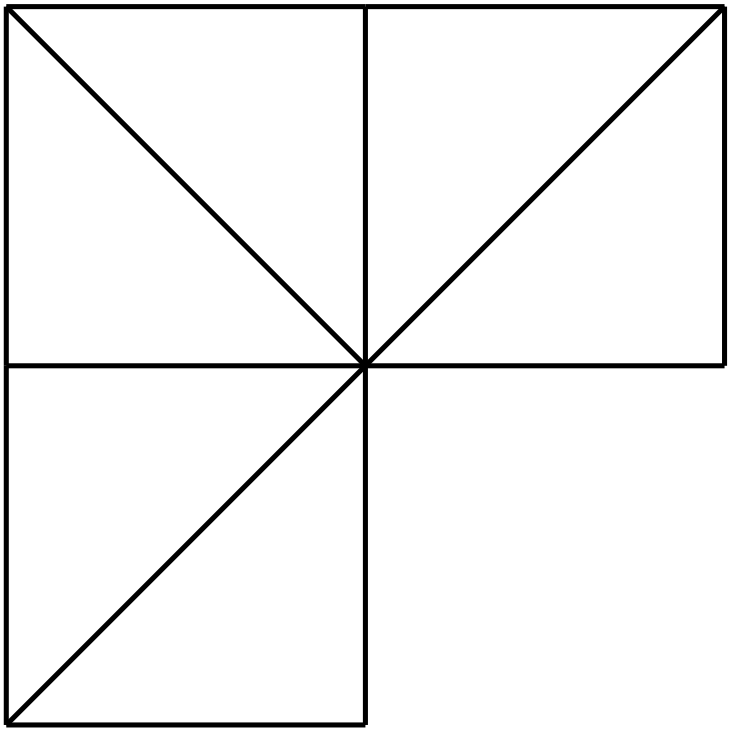


(a) Symmetric mesh of 8 triangles for a square

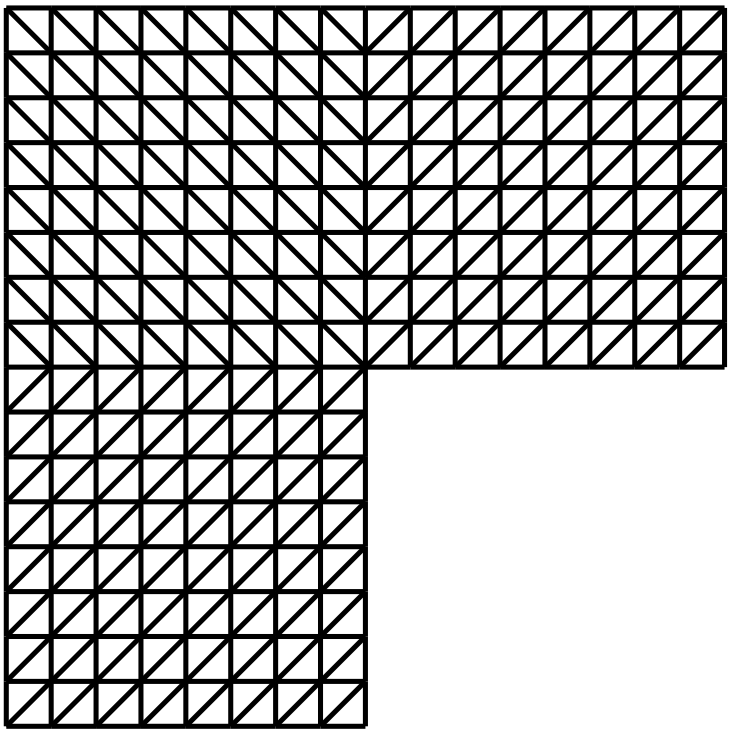


(b) The first mesh of 128 triangles used in the computations

Fig. 5.1: Symmetric meshes of a square

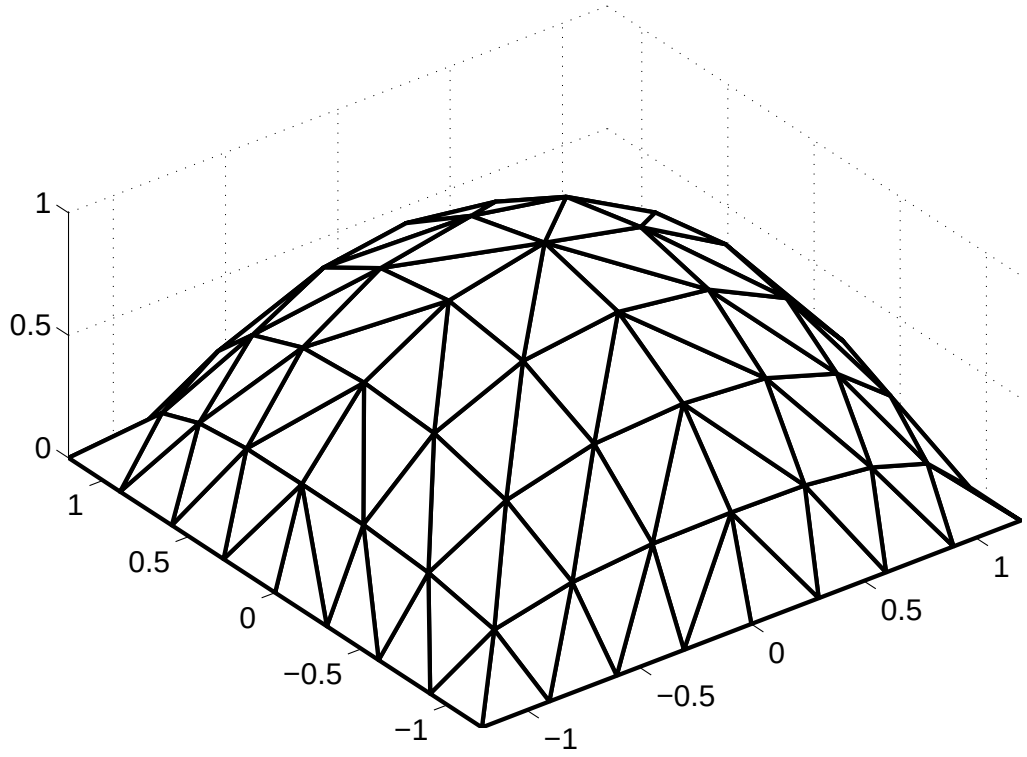


(a) Mesh of 6 triangles for a L-shape

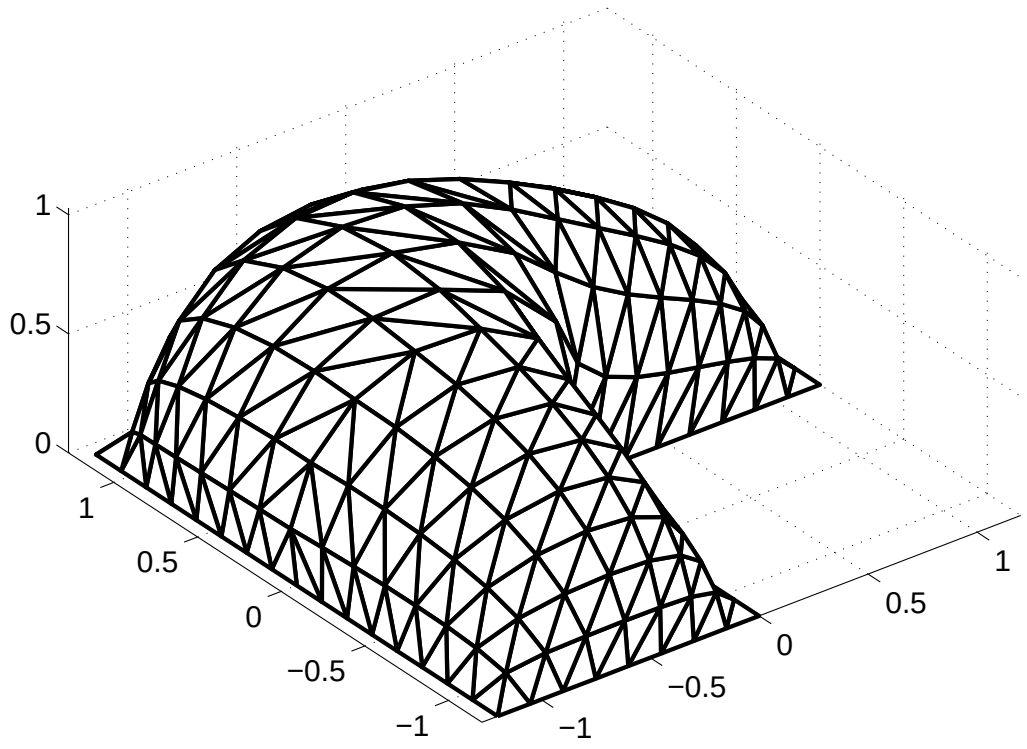


(b) The first mesh of 384 triangles used in the computations

Fig. 5.2: Symmetric meshes of a L-shape



(a) Deformed sheet with first mesh of a square when $P = 1.8$



(b) Deformed sheet with first mesh of a L-shape when $P = 3.2$

Fig. 5.3: Deformed sheets with $\max\{u_3(x_1, x_2) : (x_1, x_2) \in \Omega\}$ just above 1.

5.2 Experiments with uniform refinement

We consider uniform refinement for 3 cases when Ω is a square as follows. We take $J = J_1$ with $\Omega^* = \Omega_1^*$, we take $J = J_1$ with $\Omega^* = \Omega_2^*$ and we take $J = J_2$. When Ω is the L-shape we just consider $J = J_2$. In each case when a solution $\underline{u}_h$ with linear triangles is obtained we also compute the likely better solution $\underline{u}_h^b$ using 6-noded triangles (with the same triangular mesh) in order to compute

$$\underline{u}_h^m = \frac{1}{2}(\underline{u}_h + \underline{u}_h^b).$$

When the data for the dual problem is $\underline{u}_h$ we denote the dual solution by $\underline{\psi}_h$ and when the data is $\underline{u}_h^m$ we denote the dual solution by $\underline{\psi}_h^m$. If we take the most accurate estimates of $J(\underline{u})$ to be $J(\underline{u}_h^b)$ for the finest of the meshes used then these are given in table 5.2.1.

In tables 5.2.2, 5.2.3 and 5.2.4 we show 3 estimates of the true error and these are

$$-a(\underline{u}_h; \underline{\psi}_h), \quad -a(\underline{u}_h; \underline{\psi}_h^m) \quad \text{and} \quad J(\underline{u}_h^b) - J(\underline{u}_h). \tag{5.2.1}$$

The column "Ratio using $\underline{\psi}_h^m$" is the ratio of successive estimates using $\underline{\psi}_h^m$ from the uniform refinement. For the square domain successive steps of uniform refinement reduces the error by about 4 which is the expected rate of convergence. Comparison of the columns $-a(\underline{u}_h; \underline{\psi}_h^m)$ and $J(\underline{u}_h^b) - J(\underline{u}_h)$ show that these two estimates are very close. The easier to compute estimate $-a(\underline{u}_h; \underline{\psi}_h)$ is reasonably close for all the cases when $J = J_1$ but it appear to over estimate the error by a factor close to 2 in the case of $J = J_2$.

For the L-shape domain the regularity of the solution is not high enough to get a decrease by a factor close to 4 but it is still a bit more than 3 and again $-a(\underline{u}_h; \underline{\psi}_h^m)$ and $J_2(\underline{u}_h^b) - J_2(\underline{u}_h)$ are very close. The easier to compute estimate $-a(\underline{u}_h; \underline{\psi}_h)$ overestimates the actual error by a factor close to 2 as it did in the case when Ω is a square.

The results of all the tables demonstrates that for these test problems the prediction of the error in an estimate of a quantity of interest is good with a relatively small number of elements. A major criticism of the overall set-up is however the amount of effort to estimate the error especially for the larger values of the number of elements *ne*. In the largest case presented *ne*=24576 and when 6-noded quadratic elements are used this involves 148995 unknowns and 5105673 entries in a sparsely stored Jacobian matrix or for the matrix in the dual solution. From a practical point of view some of computation can quite easily be avoided by considering the ratios $a(\underline{u}_h; \underline{\psi}_h)/a(\underline{u}_h; \underline{\psi}_h^m)$ as the mesh size varies and we show these in table 5.2.6. In each case as *ne* varies, the ratios do not change

Tab. 5.2.1: Estimates of the QoI with the finest meshes used and 6-noded triangular elements

Domain	ne	J	QoI estimate $J(\underline{u}_h^b)$
Square, $\Omega^* = \Omega_1^*$	8192	J_1	3.272310533e-01
Square, $\Omega^* = \Omega_2^*$	8192	J_1	4.956794633e-01
Square	8192	J_2	-1.779109466e+00
L-shape	24576	J_2	-2.798976431e+00

Tab. 5.2.2: Estimates of the error $J_1(\underline{u}) - J_1(\underline{u}_h)$ for a square when $\Omega^* = \Omega_1^*$, i.e. an average of λ_3 close to the centre, with uniform refinement.

ne	$-a(\underline{u}_h;\underline{\psi}_h)$	$-a(\underline{u}_h;\underline{\psi}_h^m)$	$J_(\underline{u}_h^b) - J_1(\underline{u}_h)$	Ratio using $\underline{\psi}_h^m$
128	-1.451803e-02	-1.577692e-02	-1.565002e-02	
512	-3.926582e-03	-4.257620e-03	-4.250058e-03	3.705573e+00
2048	-1.007150e-03	-1.093677e-03	-1.093207e-03	3.892941e+00
8192	-2.537830e-04	-2.758147e-04	-2.757853e-04	3.965260e+00

Tab. 5.2.3: Estimates of the error $J_1(\underline{u}) - J_1(\underline{u}_h)$ for a square when $\Omega^* = \Omega_2^*$, i.e. an average of λ_3 close to the edge, with uniform refinement.

ne	$-a(\underline{u}_h;\underline{\psi}_h)$	$-a(\underline{u}_h;\underline{\psi}_h^m)$	$J_1(\underline{u}_h^b) - J_(\underline{u}_h)$	Ratio using $\underline{\psi}_h^m$
128	-2.152254e-02	-2.205487e-02	-2.194325e-02	
512	-6.138624e-03	-6.341243e-03	-6.335320e-03	3.478004e+00
2048	-1.608791e-03	-1.669629e-03	-1.669276e-03	3.797995e+00
8192	-4.096752e-04	-4.260730e-04	-4.260511e-04	3.918645e+00

Tab. 5.2.4: Estimates of the error $J_2(\underline{u}) - J_2(\underline{u}_h)$, i.e. the potential energy, for a square with uniform refinement.

ne	$-a(\underline{u}_h;\underline{\psi}_h)$	$-a(\underline{u}_h;\underline{\psi}_h^m)$	$J_2(\underline{u}_h^b) - J_2(\underline{u}_h)$	Ratio using $\underline{\psi}_h^m$
128	-2.528180e-01	-1.275640e-01	-1.267430e-01	
512	-6.786452e-02	-3.410450e-02	-3.403597e-02	3.740386e+00
2048	-1.736280e-02	-8.696848e-03	-8.691775e-03	3.921478e+00
8192	-4.370589e-03	-2.186427e-03	-2.186091e-03	3.977653e+00

Tab. 5.2.5: Estimates of the error $J_2(\underline{u}) - J_2(\underline{u}_h)$, i.e. the potential energy, for a L-shape with uniform refinement.

ne	$-a(\underline{u}_h;\underline{\psi}_h)$	$-a(\underline{u}_h;\underline{\psi}_h^m)$	$J_2(\underline{u}_h^b) - J_2(\underline{u}_h)$	Ratio using $\underline{\psi}_h^m$
384	-5.718256e-01	-2.806478e-01	-2.772122e-01	
1536	-1.841024e-01	-9.065824e-02	-9.010840e-02	3.095668e+00
6144	-5.699727e-02	-2.817260e-02	-2.805969e-02	3.217958e+00
24576	-1.813532e-02	-8.958059e-03	-8.919452e-03	3.144945e+00

very much. We have not tried to analyse whether or not this is generally the case but if it is then it would appear reasonable to first compute everything to get the first two rows in the tables to get a number

$$\gamma_H = \frac{a(\underline{u}_H; \underline{\psi}_H)}{a(\underline{u}_H; \underline{\psi}_H^m)}, \tag{5.2.2}$$

where H denotes the mesh size for the row considered. For larger values of ne we can then just compute $\underline{u}_h$, $\underline{\psi}_h$ and $-a(\underline{u}_h; \underline{\psi}_h)$ and take the following as our improved estimate of the error

$$\frac{-a(\underline{u}_h; \underline{\psi}_h)}{\gamma_H}. \tag{5.2.3}$$

Tab. 5.2.6: The ratio $a(\underline{u}_h; \underline{\psi}_h)/a(\underline{u}_h; \underline{\psi}_h^m)$ of the error estimates in the following 4 cases.

ne	J_1 with Ω_1^*	J_1 with Ω_2^*	J_2 (square)	ne	J_2 (L-shape)
128	9.202069e-01	9.758634e-01	1.981891e+00	384	2.037520e+00
512	9.222481e-01	9.680474e-01	1.989899e+00	1536	2.030730e+00
2048	9.208843e-01	9.635620e-01	1.996447e+00	6144	2.023146e+00
8192	9.201214e-01	9.615141e-01	1.998964e+00	24576	2.024470e+00

5.3 Non-uniform refinement with the L-shape

Experiments when Ω is a square and adaptive refinement based on the element values

$$-a(\underline{u}_h; \tilde{\underline{\psi}}_h - \tilde{\underline{\psi}}_I)_k, \quad k = 1, 2, \ldots, ne, \tag{5.3.1}$$

where $\tilde{\underline{\psi}}_h$ is either $\underline{\psi}_h$ or $\underline{\psi}_h^m$ of the dual solutions in $\bar{V}_h$ and where $\tilde{\underline{\psi}}_I \in \bar{V}_h$ is the interpolant of $\tilde{\underline{\psi}}_h$, has mostly led to uniform convergence or very close to uniform convergence as the decision as to which triangles to refine to get a more accurate solution. As a consequence we only give the results in the case when Ω is a L-shape.

In the case of the L-shape, if we take as our aim to get an estimate $J_2(\underline{u}_h)$ to be within $\epsilon = 10^{-2}$ of the true value then table 5.2.5 shows that this is only first the case when ne=24576 elements when we uniformly refine. Part of the difficulty of getting high accuracy with the L-shape is the re-entrant corner and this it helps to start with a mesh with smaller triangles near the corner and thus before the computations start we successively refine all the triangles which has $(0, 0)$ as one of the nodes. If we do this once then we get the mesh of 408 triangles shown in figure 5.4(a) and in we do this in total 4 times then we get the mesh of 480 triangles shown in figure 5.4(b). We use the mesh of 480 triangles as the first mesh in the computations. From table 5.2.1 the value

$J_2(\underline{u}_h^b) \approx -2.8$ and the accuracy of the error estimate $J_2(\underline{u}_h^b) - J_2(\underline{u}_h)$ with 480 elements is about -2.3×10^{-1}, see table 5.3.7. We need to reduce this by a factor of about 23 to get the desired accuracy.

A difficulty with using the numbers

$$\epsilon_k = -a(\underline{u}_h; \underline{\tilde{\psi}}_h - \underline{\tilde{\psi}}_I)_k, \quad k = 1, 2, \ldots, ne, \tag{5.3.2}$$

to decide which elements to refine is that they are not all of the same sign for many quantities of interest and this is the case with the L-shape. If we want to compute until

$$\left| \sum_{k=1}^{ne} \epsilon_k \right| < \epsilon \tag{5.3.3}$$

then a condition such as refining all triangles for which

$$|\epsilon_k| > \frac{\epsilon}{\mathrm{ne}} \tag{5.3.4}$$

can typically lead to all or almost all triangles being refined. Also, the outcome when we refine all the triangles is to reduce the error by a factor close to 4 at best and if we only refine some of the triangles then we are likely to reduce the error by less than 4. If, for example, two refinements are done which each reduce the error by about a factor of 2 then this needs to involve less overall computation than one uniform refinement to be worthwhile. As we need to reduce the error by about 23, we thus need at least 3 refinement steps. So far we do not have any strategy which works well enough to be clearly better than uniform refinement until we are close enough to the target accuracy. The outcome with 2 steps of uniform refinement is shown in table 5.3.7 with the error reduced to about -1.7×10^{-2} with $7680 = 16 \times 480$ elements. At this stage we should be able to get below 10^{-2} without refining all the triangles. If for the description we assume that the numbering of the triangles are such that

$$|\epsilon_1| \geq |\epsilon_2| \geq \cdots \geq |\epsilon_{ne}| \tag{5.3.5}$$

then we wish to select K, $1 \leq K \leq ne$ such that we just refine elements $1, 2, \ldots, K$. A fairly crude reasoning is that if we just refine these K triangles then error may be about

$$q_K = \left(\frac{1}{4} \sum_{k=1}^{K} \epsilon_k \right) + \sum_{k=K+1}^{ne} \epsilon_k. \tag{5.3.6}$$

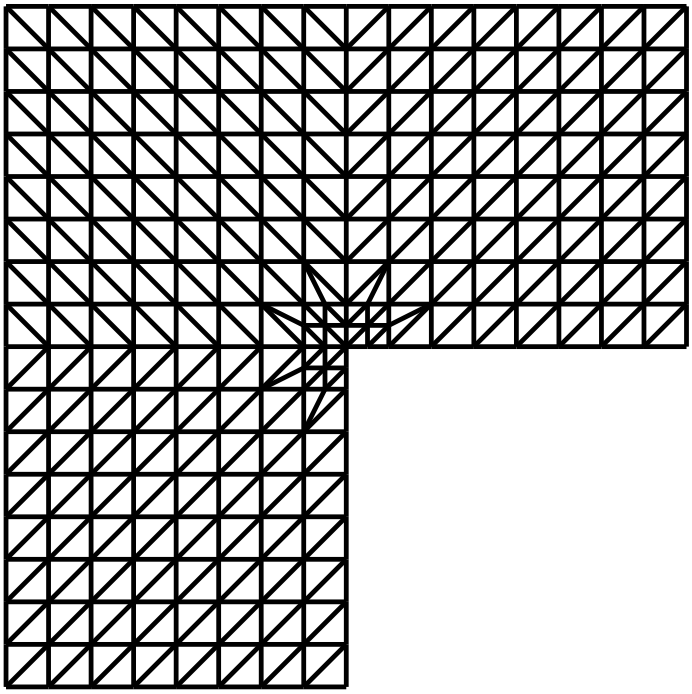
This would suggest selecting K such that

$$\epsilon < |q_{K-1}| \quad \text{and} \quad |q_K| \leq \epsilon. \tag{5.3.7}$$

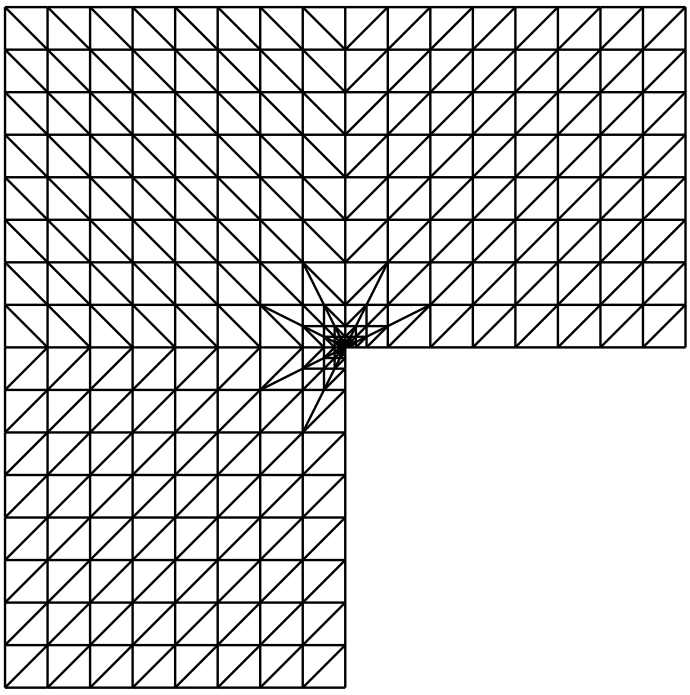
Unfortunately this prediction of what to refine to get the required accuracy has never been enough but instead refining elements $1, 2, \ldots, 2K$ seems generally to be enough and this is what is done when refining the mesh of 7680 to get the mesh of 17182 elements given in table 5.3.7. It is actually a big saving to just use 17182 elements compared with 24576 in table 5.2.5 or to use $30720 = 4 \times 7680$ if we were to uniformly refine the last mesh. A recommendation from these experiments is not to try to refine too few elements when attempting an adaptive refinement step. Also, as in the case of uniform refinement at each step, we can save a lot of computation by just using $-a(\underline{u}_h; \underline{\psi}_h)/\gamma_H$ once we have a suitable estimate for the ratio estimates γ_H as defined in (5.2.2).

Tab. 5.3.7: Non-uniform refinement in the last step with the L-shape domain and the QoI is the total potential energy $J_2(\underline{u})$. The column "Ratio" is the ratio of successive estimates of $-a(\underline{u}_h; \underline{\psi}_h^m)$ as ne is increased.

ne	$-a(\underline{u}_h; \underline{\psi}_h)$	$-a(\underline{u}_h; \underline{\psi}_h^m)$	$a(\underline{u}_h; \underline{\psi}_h)/a(\underline{u}_h; \underline{\psi}_h^m)$	$J_2(\underline{u}_h^b) - J_2(\underline{u}_h)$	Ratio
480	-4.633590e-01	-2.288500e-01	2.024728e+00	-2.268550e-01	
1920	-1.304188e-01	-6.478859e-02	2.012990e+00	-6.458486e-02	3.532
7680	-3.399703e-02	-1.695775e-02	2.004808e+00	-1.693907e-02	3.821
17182	-1.939180e-02	-9.679853e-03	2.003316e+00	-9.673313e-03	1.752



(a) Mesh of 408 triangles for the L-shape.



(b) Mesh of 480 triangles for the L-shape.

Fig. 5.4: Meshes of the L-shape with refinement about the re-entrant corner at $(0, 0)$. In the top figure the refinement is done once and it is done 4 times in the bottom figure.

6. THE AXISYMMETRIC MEMBRANE MODEL – THE QUASI-STATIC AND DYNAMIC CASES

6.1 Introduction

In this chapter we simplify the geometry to circular disks deforming under axisymmetric conditions. In the quasi-static case, the problem now is reduced to one dimension with all the unknowns depending on the radial direction r in a cylindrical polar coordinate system. This is the simplest of the cases described in this thesis. However much of the chapter is about the more difficult case when the model involves the equations of motion with quantities now depending on both r and the time t and this is what we refer to the dynamic case. In the quasi-static case we just have the displacement $\underline{u} = \underline{u}(r)$ as the unknown but in the dynamic case we have a weak form description of the problem with both the displacement $\underline{u}(r,t)$ and the velocity $\underline{v}(r,t)$ as unknowns. The weak form descriptions are described in sections 6.3 and 6.4. The remainder of the chapter is about a basic finite element scheme to approximately solve the problem described in the weak form.

The applications of the techniques are described in chapter 8.

6.2 The membrane deformation in the axisymmetric case

Here, we describe the problem of the inflation of a circular membrane which initially has uniform thickness h_0, which is clamped around its circumference. First we give the equations for a general three-dimensional body. Then we show how the axisymmetry with the membrane properties reduce the problem to one space dimension in which the unknowns depend only on the radial dimension in a cylindrical polar coordinate system.

The region of the undeformed body of an axisymmetric shape, with respect to cylindrical

polars, can be represented by

$$\{(r, \theta, x_3) : 0 \leq r < 1,\ -\pi < \theta \leq \pi,\ |x_3| < h_0/2\}\,.$$

This is deformed by a time-dependent pressure $P(t)$ applied to the lower side of the circular sheet. We let Ω be the undeformed mid-surface and, for this case, this is just $\Omega = [0, 1)$ with respect to r. The deformation of the midsurface now simplifies to

$$(r, \theta, 0) \to (r + u_1, \theta, u_3) =: \underline{w}$$

where $u_1 = u_1(r)$ is the radial displacement and $u_3 = u_3(r)$ is the vertical displacement. As before, $\underline{w}$ denotes the deformed midsurface for the axisymmetric case. In everything that follows in this section, we consider the membrane model of how the sheet deforms with the details essentially being the axisymmetric version of what was given in section 3.4. At the start of the process when the sheet is flat the membrane is uniformly pre-stretched and then clamped at the edge $r = 1$ so that $\underline{u}(1, t)$ does not vary with t, i.e.

$$u_1(r, 0) = u_1(1, 0)r, \quad u_3(r, 0) = 0 \quad \text{and} \quad \underline{u}(1, t) = \underline{u}(1, 0) \text{ for all } t > 0.$$

The density of the material is denoted by ρ and we assume that the body is composed of homogeneous, isotropic, incompressible, hyperelastic material. Thus, in particular, if ρ_0 is the density initially then $\rho = \rho_0$ is the density throughout the deformation.

In the following description the symbol $'$ is used to denote differentiation with respect to r and we now let $\underline{e}_1$ and $\underline{e}_3$ denote the base vectors in the radial and vertical directions respectively. The base vector $\underline{e}_2 = \underline{e}_\theta = \underline{e}_3 \times \underline{e}_1$ denotes the base vector in the θ direction.

6.2.1 The principal stretches for the axisymmetric membrane case

The **membrane deformation gradient** evaluated on the midsurface, by using cylindrical polars, is given by

$$\mathbf{F} = \begin{pmatrix} 1 + u_1' & 0 \\ 0 & 1 + \frac{u_1}{r} \\ u_3' & 0 \end{pmatrix}.$$

We also define **the right Cauchy Green deformation tensor** by

$$\mathbf{C} = \mathbf{F}^T\mathbf{F}$$

and let

$$\lambda_1^2 = (1+u_1')^2 + u_3'^2, \quad \lambda_2 = 1 + \frac{u_1}{r}.$$

As before, λ_1 and λ_2 denote the principal stretches which correspond to the directions

$$\underline{t} := \underline{b}_1 = \left(\frac{1+u_1'}{\lambda_1}\right)\underline{e}_1 + \left(\frac{u_3'}{\lambda_1}\right)\underline{e}_3 \quad \text{and} \quad \underline{b}_2 = \underline{e}_\theta,$$

with the unit normal $\underline{n}$ being

$$\underline{n} := \underline{b}_3 = -\left(\frac{u_3'}{\lambda_1}\right)\underline{e}_1 + \left(\frac{1+u_1'}{\lambda_1}\right)\underline{e}_3, \tag{6.2.1}$$

where $\underline{b}_1, \underline{b}_2, \underline{b}_3$ denote the three principal directions with respect to the deformed configuration.

Other useful relations to give are the membrane deformation gradient in 3D and its polar decomposition, which are given by

$$\mathbf{F}_{3D} = \mathbf{R}\mathbf{U}, \quad \mathbf{R} = (\underline{t}, \underline{e}_\theta, \underline{n}), \quad \mathbf{U} = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix},$$

where $\lambda_3 = 1/(\lambda_1\lambda_2)$ which denotes the **thickness stretch ratio** (assuming incompressibility).

It is worth commenting on what happens to some of these terms as $r \to 0$ with $r = 0$ being the pole of the circular disk. Firstly, for λ_2 to have a finite limit as $r \to 0$ we require that $u_1(0) = 0$ and this gives

$$\lambda_2(0) = \lim_{r\to 0} \lambda_2(r) = 1 + u_1'(0). \tag{6.2.2}$$

The curvatures of the deformed surface are concerned with changes in the unit normal $\underline{n}$ as we move on the deformed surface. If for the moment we consider $\underline{n} = \underline{n}(r,\theta)$ then

$$\frac{\partial \underline{e}_1}{\partial \theta} = \underline{e}_2 \quad \text{which gives} \quad \frac{\partial \underline{n}}{\partial \theta} = -\frac{u_3'}{\lambda_1}\underline{e}_2$$

and when we consider the change relative to the change of the positions of points on the surface we have

$$\frac{\underline{n}(r,\theta+\Delta\theta) - \underline{n}(r,\theta)}{(r+u_1(r))\Delta\theta} \to \frac{-u_3'(r)}{(r+u_1(r))\lambda_1}\underline{e}_2 \quad \text{as } \Delta\theta \to 0. \tag{6.2.3}$$

For the right hand side to have a finite limit as $r \to 0$ requires that $u_3'(0) = 0$ and as a consequence

$$\lambda_1(0) = \lambda_2(0) = 1 + u_1'(0). \tag{6.2.4}$$

6.2.2 The principal stresses for the axisymmetric membrane case

In a membrane approximation of how a thin sheet deforms we assume that in the normal direction

$$\boldsymbol{\sigma}\underline{n} = \underline{0}$$

where $\boldsymbol{\sigma}$ denotes the **Cauchy stress**. With the axisymmetric deformation we also have that in the $\underline{e}_\theta$ direction

$$\boldsymbol{\sigma}\underline{e}_\theta = \sigma_2\underline{e}_\theta$$

where σ_2 is a principal stress. Finally the other direction of principal stress is $\underline{t}$ with

$$\boldsymbol{\sigma}\underline{t} = \sigma_1\underline{t}$$

where σ_1 is the other principal stress. Hence since $\mathbf{R} = (\underline{t}, \underline{e}_\theta, \underline{n})$, by using the directions of principal stresses we have the following

$$\mathbf{R}^T\boldsymbol{\sigma}\mathbf{R} = \begin{pmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{6.2.5}$$

This is the axisymmetric case of (3.4.10).

Next we define the **nominal stress $\boldsymbol{\Pi}_{3D}$** and **the first Piola stress $\boldsymbol{\Pi}_{3D}^T$** for a three dimensional deformation, which have the following representation respectively.

$$\boldsymbol{\Pi}_{3D} = (\det \mathbf{F}_{3D})\mathbf{F}_{3D}^{-1}\boldsymbol{\sigma},$$
$$\boldsymbol{\Pi}_{3D}^T = \boldsymbol{\sigma}(\det \mathbf{F}_{3D})\mathbf{F}_{3D}^{-T} = \boldsymbol{\sigma}\mathbf{F}_{3D}^{-T} = \sigma\mathbf{R}\mathbf{U}^{-1} = \mathbf{R}(\mathbf{R}^T\boldsymbol{\sigma}\mathbf{R})\mathbf{U}^{-1} = \left(\frac{\sigma_1}{\lambda_1}\underline{t}, \frac{\sigma_2}{\lambda_2}\underline{e}_\theta, \underline{0}\right),$$

where $\det \mathbf{F}_{3D} = \lambda_1\lambda_2\lambda_3 = 1$ assuming incompressibility and $\mathbf{R}\mathbf{R}^T = \mathbf{R}^T\mathbf{R} = \mathbf{I}$ since $\mathbf{R}$ is orthogonal matrix. For the membrane simplification for the axisymmetric case we get the following

$$\boldsymbol{\Pi}^T = \left(\frac{\sigma_1}{\lambda_1}\underline{t}, \frac{\sigma_2}{\lambda_2}\underline{e}_\theta\right) = \begin{pmatrix} \frac{(1+u_1')\sigma_1}{\lambda_1^2} & 0 \\ 0 & \frac{\sigma_2}{\lambda_2} \\ \frac{u_3'\sigma_1}{\lambda_1^2} & 0 \end{pmatrix} \tag{6.2.6}$$

which represents **the membrane first Piola stress** for the axisymmetric case.

To complete the mathematical description, we need to describe the constitutive relationship between stress and stretch that, in the case of a hyperelastic, incompressible and isotropic material, as we assumed for this model, can be expressed as follows

$$\Pi^T = \frac{\partial W}{\partial F} \quad \text{and} \quad \sigma_i = \lambda_i \frac{\partial W}{\partial \lambda_i} \quad i = 1, 2,$$

as we described in section 3.5, with W being the strain energy function.

6.3 The weak form for the quasi-static case

For the simpler quasi-static case the dependence on time t is only through the time-dependent pressure loading. Therefore for fixed time a weak form is given by

$$A_Q(t)(\underline{u}; \underline{\psi}) = 0, \quad \forall \underline{\psi} \in V \tag{6.3.1}$$

where $V = H_0^1(\Omega)$ and $t \in [0, T]$ with T being the final time. As a consequence of what is described in Chapter 3 the quasi-static problem in weak form, by using cylindrical polars, involves the following.
For fixed time $t \in [0, T]$, find $\underline{u} \in V$ such that

$$A_Q(t)(\underline{u}; \underline{\psi}) = h_0 a_1(\underline{u}; \underline{\psi}) - P(t) a_2(\underline{u}; \underline{\psi}) \quad \forall \underline{\psi} \in V$$

where

$$a_1(\underline{u}; \underline{\psi}) = \int_0^1 \mathbf{\Pi}^T : \nabla \underline{\psi}\, r \mathrm{d}r, \tag{6.3.2}$$

$$a_2(\underline{u}; \underline{\psi}) = \int_0^1 \lambda_1 \lambda_2 \underline{n} \cdot \underline{\psi}\, r \mathrm{d}r, \tag{6.3.3}$$

where for the second term we used $\mathbf{F}^{-T} \underline{e}_3 = (1/\lambda)\underline{n} = \lambda_1 \lambda_2 \underline{n}$.

There are other ways in which we can write the integrands in (6.3.2) and (6.3.3) which we use in the computational schemes and these are described next.

Expressions for the a_1 term

For the first term $a_1(\underline{u};\underline{\psi})$ we have the following. With

$$\mathbf{\Pi}^T = \begin{pmatrix} \frac{(1+u_1')\sigma_1}{\lambda_1^2} & 0 \\ 0 & \frac{\sigma_2}{\lambda_2} \\ \frac{u_3'\sigma_1}{\lambda_1^2} & 0 \end{pmatrix} \quad \text{and} \quad \nabla\underline{\psi} = \begin{pmatrix} \psi_1' & 0 \\ 0 & \frac{\psi_1}{r} \\ \psi_3' & 0 \end{pmatrix}, \tag{6.3.4}$$

we get the following expression

$$\begin{aligned} \mathbf{\Pi}^T : \nabla\underline{\psi} &= \Pi_{11}\psi_1' + \Pi_{13}\psi_3' + \Pi_{22}\frac{\psi_1}{r} \\ &= \frac{\sigma_1}{\lambda_1^2}((1+u_1')\psi_1' + u_3'\psi_3') + \frac{\sigma_2}{\lambda_2}\frac{\psi_1}{r}. \end{aligned}$$

Further, if we use the notation

$$W_1 := \frac{\partial W}{\partial \lambda_1}, \quad \text{and} \quad W_2 := \frac{\partial W}{\partial \lambda_2},$$

which was used in (3.6.27), then our expression for $a_1(\cdot;\cdot)$ becomes

$$\mathbf{\Pi}^T : \nabla\underline{\psi} = \frac{W_1}{\lambda_1}((1+u_1')\psi_1' + u_3'\psi_3') + W_2\frac{\psi_1}{r}. \tag{6.3.5}$$

Thus

$$a_1(\underline{u};\underline{\psi}) = \int_0^1 \left(\frac{W_1}{\lambda_1}((1+u_1')\psi_1' + u_3'\psi_3') + W_2\frac{\psi_1}{r}\right) r\mathrm{d}r. \tag{6.3.6}$$

Expressions for the a_2 term

For the second term $a_2(\underline{u};\underline{\psi})$ we have the following. By using (6.2.1) we get

$$\lambda_1\lambda_2\underline{n}\cdot\underline{\psi} = \lambda_2\left(-u_3'\psi_1 + (1+u_1')\psi_3\right), \tag{6.3.7}$$

and hence

$$\begin{aligned} a_2(\underline{u};\underline{\psi}) &= \int_0^1 \lambda_2\left(-u_3'\psi_1 + (1+u_1')\psi_3\right) r\mathrm{d}r, \\ &= \int_0^1 (r+u_1)(-u_3'\psi_1 + (1+u_1')\psi_3)dr. \end{aligned} \tag{6.3.8}$$

As in section 3.4.2 we refer to this term $a_2(\cdot;\cdot)$ as the pressure term of the weak form. Now in section 3.4.2 we gave a longer way of writing this enabled us to show that we had symmetry in the matrices involved in a Newton iteration and in a related dual problem. The corresponding version of this in this axisymmetric case is

$$\begin{aligned} a_2(\underline{u};\underline{\psi}) &= \frac{1}{3}\int_0^1 (r+u_1)((1+u_1')\psi_3 - (r+u_1)\psi_3')dr \\ &+ \frac{1}{3}\int_0^1 (u_3((r+u_1)\psi_1' + (1+u_1')\psi_1) - 2(r+u_1)u_3'\psi_1)dr. \end{aligned} \tag{6.3.9}$$

To verify that (6.3.7) and (6.3.9) are the same can be done as follows.

As it was described in chapter 3, the cartesian form of the long expression of the pressure term is given by

$$a_2(\underline{u};\underline{\psi}) = \frac{1}{3}\iint_\Omega \left(\underline{\psi}\cdot\left(\frac{\partial \underline{w}}{\partial x_1}\times\frac{\partial \underline{w}}{\partial x_2}\right) + \underline{w}\cdot\left(\frac{\partial \underline{\psi}}{\partial x_1}\times\frac{\partial \underline{w}}{\partial x_2}\right) + \underline{w}\cdot\left(\frac{\partial \underline{w}}{\partial x_1}\times\frac{\partial \underline{\psi}}{\partial x_2}\right)\right) dx_1 dx_2. \tag{6.3.10}$$

Now, by using a cylindrical polar coordinate system, the above expression becomes

$$a_2(\underline{u};\underline{\psi}) = \frac{1}{3}\int_0^1 \left(\underline{\psi}\cdot\left(\frac{\partial \underline{w}}{\partial r}\times\frac{1}{r}\frac{\partial \underline{w}}{\partial \theta}\right) + \underline{w}\cdot\left(\frac{\partial \underline{\psi}}{\partial r}\times\frac{1}{r}\frac{\partial \underline{w}}{\partial \theta}\right) + \underline{w}\cdot\left(\frac{\partial \underline{w}}{\partial r}\times\frac{1}{r}\frac{\partial \underline{\psi}}{\partial \theta}\right)\right) r dr. \tag{6.3.11}$$

Then, by using cylindrical polars we get the following expressions.

$$\underline{w} = (r+u_1)\underline{e}_1(\theta) + u_3\underline{e}_3 \qquad \underline{\psi} = \psi_1\underline{e}_1(\theta) + \psi_3\underline{e}_3, \tag{6.3.12}$$

$$\frac{1}{r}\frac{\partial \underline{w}}{\partial \theta} = \left(1+\frac{u_1}{r}\right)\underline{e}_\theta \qquad \frac{1}{r}\frac{\partial \underline{\psi}}{\partial \theta} = \frac{\psi_1}{r}\underline{e}_\theta, \tag{6.3.13}$$

$$\frac{\partial \underline{w}}{\partial r} = (1+u_1')\underline{e}_1 + u_3'\underline{e}_3 \qquad \frac{\partial \underline{\psi}}{\partial r} = \psi_1'\underline{e}_1 + \psi_3'\underline{e}_3. \tag{6.3.14}$$

By using the above expressions, we compute the cross products of the integral (6.3.11). Thus we get the following quantities.

$$\frac{\partial \underline{w}}{\partial r}\times\frac{1}{r}\frac{\partial \underline{w}}{\partial \theta} = \left(1+\frac{u_1}{r}\right)\left((1+u_1')\underline{e}_3 - u_3'\underline{e}_1\right), \tag{6.3.15}$$

$$\frac{\partial \underline{\psi}}{\partial r}\times\frac{1}{r}\frac{\partial \underline{w}}{\partial \theta} = \left(1+\frac{u_1}{r}\right)(\psi_1'\underline{e}_3 - \psi_3'\underline{e}_1), \tag{6.3.16}$$

$$\frac{\partial \underline{w}}{\partial r}\times\frac{1}{r}\frac{\partial \underline{\psi}}{\partial \theta} = \frac{\psi_1}{r}\left((1+u_1')\underline{e}_3 - u_3'\underline{e}_1\right). \tag{6.3.17}$$

Finally, by using the above quantities and the expressions (6.3.12) we compute the dot products of the integral (6.3.11) which give us the desired longer version for the pressure

term as

$$\begin{aligned} a_2(\underline{u};\underline{\psi}) &= \frac{1}{3}\int_0^1 ((\psi_1\underline{e}_1(\theta)+\psi_3\underline{e}_3)\cdot\left(1+\frac{u_1}{r}\right)((1+u_1')\underline{e}_3-u_3'\underline{e}_1) \\ &+ ((r+u_1)\underline{e}_1(\theta)+u_3\underline{e}_3)\cdot\left(1+\frac{u_1}{r}\right)(\psi_1'\underline{e}_3-\psi_3'\underline{e}_1) \\ &+ ((r+u_1)\underline{e}_1(\theta)+u_3\underline{e}_3)\cdot\frac{\psi_1}{r}((1+u_1')\underline{e}_3-u_3'\underline{e}_1))rdr \\ &= \frac{1}{3}\int_0^1 (r+u_1)((1+u_1')\psi_3-\psi_1u_3') \\ &+ (r+u_1)(u_3\psi_1'-(r+u_1)\psi_3')+\psi_1((1+u_1')u_3-(r+u_1)u_3')dr, \end{aligned}$$

which after some rearrangements we get the desired quantity (6.3.9).

6.4 The weak form for the dynamic case

In this section we represent the equations for the dynamic case, which are described by the full equations of motion. In a PDE form and for a general 3D solid this can be written as the following system

$$\rho_0\dot{\underline{v}} = \mathrm{Div}\Pi_{3D} \tag{6.4.1}$$

$$\underline{v} = \dot{\underline{u}}, \tag{6.4.2}$$

where ρ_0 represents the undeformed density, $\underline{v}$ the velocity and $\dot{\underline{v}}$ is the acceleration. Here Div gives the divergence operation with partial derivatives with respect to the undeformed configuration. The complete description also requires initial conditions on $\underline{u}$ and $\underline{v}$. Our aim next is write the membrane version of the system (6.4.1) and (6.4.2) in a weak form. As the equations (6.4.1) and (6.4.2) hold over a region in space and time this can be written in the form

$$\Omega_{3D}\times[0,T]$$

where Ω_{3D} denotes the spatial domain and where $[0,T]$ is the time interval with T denoting the final time. Now, in order to get the weak form for the dynamic case, we take both the displacement $\underline{u}$ and the velocity $\underline{v}$ as unknowns, since we consider the full equations of motion, and we impose the relation $\underline{v}=\dot{\underline{u}}$ weakly. Then, by multiplying the PDE (6.4.1) and (6.4.2) with appropriate test vectors $\underline{\psi}$ and $\underline{\theta}$, integrating over space and time and taking the membrane approximation property into account we get the following.

With Ω denoting the undeformed mid-surface we introduce the space-time region as

$$Q = \Omega \times [0, T], \tag{6.4.3}$$

and we require that

$$a(\underline{u}; \underline{\psi})_Q + \rho h_0(\dot{\underline{v}}, \underline{\psi})_Q + \rho h_0(\dot{\underline{u}} - \underline{v}, \underline{\theta})_Q = \underline{0}, \tag{6.4.4}$$

where

$$a(\underline{u}; \underline{\psi})_Q := h_0 \int_0^T \int_0^1 a_1(\underline{u}; \underline{\psi})\ r \mathrm{d}r\, \mathrm{d}t - \int_0^T P(t) \int_0^1 a_2(\underline{u}; \underline{\psi})\ r \mathrm{d}r\, \mathrm{d}t \tag{6.4.5}$$

for $0 \leq t \leq T$, with as before h_0 being the initial thickness of the membrane. Next, we consider the initial conditions, i.e. the conditions at time $t = 0$, where we let the initial displacement be denoted by $\underline{u}^0$ and we let the initial velocity be denoted $\underline{v}^0$. For this case, we define the following

$$A\left(\begin{pmatrix}\underline{u}\\ \underline{v}\end{pmatrix}; \begin{pmatrix}\underline{\psi}\\ \underline{\theta}\end{pmatrix}\right) = a(\underline{u}; \underline{\psi})_Q + \rho h_0(\dot{\underline{v}}, \underline{\psi})_Q \tag{6.4.6}$$

$$+ \rho h_0(\dot{\underline{u}} - \underline{v}, \underline{\theta})_Q + \rho h_0(\underline{u}^0, \underline{\theta})_\Omega + \rho h_0(\underline{v}^0, \underline{\psi})_\Omega \tag{6.4.7}$$

$$\mathcal{F}\left(\begin{pmatrix}\underline{\psi}\\ \underline{\theta}\end{pmatrix}\right) = \rho h_0(\underline{u}^0, \underline{\theta})_\Omega + \rho h_0(\underline{v}^0, \underline{\psi})_\Omega. \tag{6.4.8}$$

The weak form involves finding $\underline{u}$ and $\underline{v}$ such that

$$A\left(\begin{pmatrix}\underline{u}\\ \underline{v}\end{pmatrix}; \begin{pmatrix}\underline{\psi}\\ \underline{\theta}\end{pmatrix}\right) = \mathcal{F}\left(\begin{pmatrix}\underline{\psi}\\ \underline{\theta}\end{pmatrix}\right) \quad \forall \text{ appropriate } \begin{pmatrix}\underline{\psi}\\ \underline{\theta}\end{pmatrix}. \tag{6.4.9}$$

Remarks

- We require $u_1(0, t)$, $u_1(1, t)$ and $u_3(1, t)$ be fixed and as a consequence the associated test functions $\underline{\psi}$ satisfy the conditions $\psi_1(0, t) = \psi_1(1, t) = \psi_3(1, t) = 0$. These have to do with the axisymmetric condition associated with the pole where $r = 0$ and the Dirichlet boundary conditions where $r = 1$.
- At the start, when the pressure is zero ($t = 0$), there is no difference between the quasi-static case and the dynamic case. We start with a prestretch which fixes the boundary value $u_1(1, 0) > 0$ and is such that $u_1(r, 0) = u_1(1, 0)r$, $0 \leq r \leq 1$ and for the initial velocity we take $\underline{v}(r, 0) = \underline{0}$.

6.5 The finite element method in the quasi-static case

To approximately solve (6.3.1) by the finite element method is straightforward as we just have a one dimensional problem. If we let $0 = r_0 < r_1 < \cdots < r_{ne} = 1$ denote a mesh of $0 \leq r \leq 1$ and we use piecewise polynomials of degree p on each element $[r_k, r_{k+1}]$ then, as was described in section 2.4.2, our approximation to $u_1(r)$ and $u_3(r)$ are of the form

$$(\underline{u}_h(r(s)))_i = \sum_{j=0}^{p} c_j^i b_j(s), \quad i = 1 \text{ and } 3, \tag{6.5.1}$$

where

$$r(s) = \frac{r_k + r_{k+1}}{2} + \left(\frac{r_{k+1} - r_k}{2}\right) s, \quad -1 \leq s \leq 1, \tag{6.5.2}$$

and where $b_0, \ldots, b_p$ are suitable basis functions as described in section 2.4.2. If we let $\phi_1, \ldots, \phi_n$ denote the piecewise polynomial basis function arising from considering all the elements $[r_0, r_1]$, $\ldots$, $[r_{ne-1}, r_{ne}]$ then we can write the approximations in the form

$$(\underline{u}_h(r))_1 = \sum_{i=1}^{n} d_{2i-1}\phi_i(r), \tag{6.5.3}$$

$$(\underline{u}_h(r))_3 = \sum_{i=1}^{n} d_{2i}\phi_i(r). \tag{6.5.4}$$

With degree p polynomials on each of the ne elements this gives $n = p(ne) + 1$. When we consider both components this involves $2n$ parameters which we can collect as the column vector

$$\underline{d} = (d_1, d_2, \ldots, d_{2n-1}, d_{2n})^T.$$

For each function $\phi_i(r)$ we can associate the two vector valued functions

$$\underline{H}_{2i-1}(r) = \begin{pmatrix} \phi_i(r) \\ 0 \end{pmatrix}, \quad \underline{H}_{2i}(r) = \begin{pmatrix} 0 \\ \phi_i(r) \end{pmatrix}, \quad i = 1, \ldots, n \tag{6.5.5}$$

so that

$$\underline{u}_h(r) = \sum_{i=1}^{2n} d_i \underline{H}_i(r). \tag{6.5.6}$$

As $(\underline{u}_h(0))_1 = 0$ and $\underline{u}_h(1)$ is given, it follows that 3 of the entries of $\underline{d}$ are known. The other $2n - 3$ parameters are determined by solving nonlinear equations arising from

$$A_Q(t)(\underline{u}_h; \underline{\psi}) = 0, \quad \forall \underline{\psi} \in V_h \tag{6.5.7}$$

where V_h is the span of $2n-3$ of the functions in (6.5.5). The nonlinear equations are solved by a Newton iteration with the Jacobian matrix being banded with a bandwidth of just $2p+1$ which is a consequence of each element Jacobian contribution being of size $(2p+2)\times(2p+2)$. The total amount of computational resources to get $\underline{u}_h$ is thus modest compared to the non-axisymmetric case described in previous chapters and the dynamic problem described in later sections.

6.6 A basic finite element method in the dynamic case

Compared to the description in the quasi-static case, there is much more detail needed to describe how to use the finite element method for the dynamic case. We do so here in the case of one of the simpler ways of dealing with how the approximations vary in time which was used in [28]. It is useful to have this described first before extensions are given in sections 7.6 with higher degree polynomials used in the time domain. In some places in this section we comment on how some of the detail here which is generalised in section 7.6.

6.6.1 The mesh and the parameters

To approximately solve (6.4.9) we need a space-time mesh of Q. In the time direction $0 \le t \le T$ we take time levels $0 = t_0 < t_1 < \cdots < t_N = T$ and at each time level we have a mesh of the space interval $0 \le r \le 1$. We choose to take a mesh which is fixed in time, i.e. we use the same space mesh at each time level t_j. For notation we let

$$u_1^j(r) := u_1(r,t_j), \quad u_3^j(r) := u_3(r,t_j), \tag{6.6.1}$$

$$v_1^j(r) := v_1(r,t_j), \quad v_3^j(r) := v_3(r,t_j) \tag{6.6.2}$$

denote approximations at time t_j to the components of the displacement and the velocity. (We no longer use subscript h here for the approximation as this makes the notation too messy.) To describe each of these functions, we need piecewise polynomial finite element

basis functions which we denote by $\phi_1(r), \ldots, \phi_n(r)$ such that

$$u_1^j(r) = \sum_{i=1}^{n} u_{1,i}^j \phi_i(r), \tag{6.6.3}$$

$$u_3^j(r) = \sum_{i=1}^{n} u_{3,i}^j \phi_i(r), \tag{6.6.4}$$

$$v_1^j(r) = \sum_{i=1}^{n} v_{1,i}^j \phi_i(r), \tag{6.6.5}$$

$$v_3^j(r) = \sum_{i=1}^{n} v_{3,i}^j \phi_i(r) \tag{6.6.6}$$

and this corresponds what was done in (6.5.3) and (6.5.4). When we refer to the displacement vector and the velocity vector at time t_j we mean

$$\underline{u}^j(r) := \begin{pmatrix} u_1^j(r) \\ u_3^j(r) \end{pmatrix}, \quad \underline{v}^j(r) := \begin{pmatrix} v_1^j(r) \\ v_3^j(r) \end{pmatrix}. \tag{6.6.7}$$

For each of these vectors there are $2n$ parameters and, as in the quasi-static case. We collect these as column vectors

$$\underline{c}^j = \left(u_{1,1}^j, u_{3,1}^j, \ldots, u_{1,n}^j, u_{3,n}^j\right)^T, \tag{6.6.8}$$

$$\underline{b}^j = \left(v_{1,1}^j, v_{3,1}^j, \ldots, v_{1,n}^j, v_{3,n}^j\right)^T. \tag{6.6.9}$$

To indicate with more compact notation the dependence of the vectors $\underline{u}^j(r)$ and $\underline{v}^j(r)$ on $\underline{c}^j$ and $\underline{b}^j$ respectively we associate with each function $\phi_i(r)$ the two vector valued functions

$$\underline{H}_{2i-1}(r) = \begin{pmatrix} \phi_i(r) \\ 0 \end{pmatrix} \quad \text{and} \quad \underline{H}_{2i}(r) = \begin{pmatrix} 0 \\ \phi_i(r) \end{pmatrix}, \tag{6.6.10}$$

as we did in the quasi-static case, so that

$$\underline{u}^j(r) = \sum_{i=1}^{2n} c_i^j \underline{H}_i(r), \tag{6.6.11}$$

$$\underline{v}^j(r) = \sum_{i=1}^{2n} b_i^j \underline{H}_i(r). \tag{6.6.12}$$

Between time levels, i.e. $t_{j-1} < t < t_j$ we define the approximations as

$$\underline{u}(r,t) = \left(\frac{t_j - t}{t_j - t_{j-1}}\right)\underline{u}^{j-1}(r) + \left(\frac{t - t_{j-1}}{t_j - t_{j-1}}\right)\underline{u}^{j}(r), \tag{6.6.13}$$

$$\underline{v}(r,t) = \left(\frac{t_j - t}{t_j - t_{j-1}}\right)\underline{v}^{j-1}(r) + \left(\frac{t - t_{j-1}}{t_j - t_{j-1}}\right)\underline{v}^{j}(r). \tag{6.6.14}$$

It is this degree 1 polynomial in t behaviour which distinguishes what we mean by the basic or standard approach in this thesis and it is this part that we replace by higher degree polynomials in t in a time interval $[t_{j-1}, t_j]$ in section 7.6.

6.6.2 The discrete nonlinear system to satisfy

Assuming that we start with a prestretch, as previously described, and we start with a velocity of zero we have

$$\underline{u}^0(r) = \begin{pmatrix} (\text{const})r \\ 0 \end{pmatrix}, \quad \underline{v}^0(r) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad 0 \le r \le 1 \tag{6.6.15}$$

and this determines $\underline{c}^0$ and $\underline{b}^0$ (i.e. in particular $\underline{b}^0 = \underline{0}$).

Once a solution is known at time t_{j-1} we know $\underline{c}^{j-1}$ and $\underline{b}^{j-1}$. The discrete version of (6.4.9) to determine $\underline{c}^j$ and $\underline{b}^j$ depends on which test vectors $\underline{\psi}$ and $\underline{\theta}$ are used at this stage. In the case of $\underline{\psi}$ we take the vectors

$$\underline{\psi}_k(r,t) = \begin{cases} \underline{H}_k(r), & t_{j-1} < t < t_j, \\ 0, & \text{otherwise}, \end{cases} \tag{6.6.16}$$

and similarly in the case of $\underline{\theta}$ we take

$$\underline{\theta}_k(r,t) = \begin{cases} \underline{H}_k(r), & t_{j-1} < t < t_j, \\ 0, & \text{otherwise}. \end{cases} \tag{6.6.17}$$

In both cases we take all the values of $k = 1, \ldots, 2n$ except those corresponding to known

boundary values at $r = 0$ or at $r = 1$. We separately have to satisfy

$$A\left(\begin{pmatrix}\underline{u}\\ \underline{v}\end{pmatrix};\begin{pmatrix}\underline{\psi}_k\\ \underline{0}\end{pmatrix}\right) = 0 \quad \forall \text{ appropriate } k, \tag{6.6.18}$$

$$A\left(\begin{pmatrix}\underline{u}\\ \underline{v}\end{pmatrix};\begin{pmatrix}\underline{0}\\ \underline{\theta}_k\end{pmatrix}\right) = 0 \quad \forall \text{ appropriate } k. \tag{6.6.19}$$

Note that the test functions are constant in time t in the time interval. This is something we generalise in section 7.6 when we have more unknowns related to the time interval $[t_{j-1}, t_j]$ and we need more test vectors so that the number of equations matches the number of unknowns.

Terms that involve only $\underline{\theta}$-functions

We consider the "$\underline{\theta}_k$ equations" given in (6.6.19) first as this gives a connection between the vectors $\underline{b}^j$ and $\underline{c}^j$ (or equivalently between the displacement and the velocity) which can be used in (6.6.18) to get a nonlinear system involving $\underline{c}^j$ only. The details are as follows.

The "$\underline{\theta}_k$ equations" in terms of the functions are that

$$(\dot{\underline{u}} - \underline{v}, \underline{\theta}_k)_Q = \int_{t_{j-1}}^{t_i}\int_0^1 (\dot{\underline{u}}(r,t) - \underline{v}(r,t)) \cdot \underline{H}_k(r)\; r\mathrm{d}r\mathrm{d}t = 0, \quad \forall \text{ appropriate } k. \tag{6.6.20}$$

For $t_{j-1} < t < t_j$ the term $\dot{\underline{u}}$ does not vary with t and the velocity varies with t as described in (6.6.14) and we can do both of the integrals in time exactly to give

$$(t_j - t_{j-1})\int_0^1 \left(\frac{\underline{u}^j(r) - \underline{u}^{j-1}(r)}{t_j - t_{j-1}} - \frac{1}{2}\left(\underline{v}^j(r) + \underline{v}^{j-1}(r)\right)\right) \cdot \underline{H}_k(r)\; r\mathrm{d}r = 0. \tag{6.6.21}$$

In terms of the parameters which define the displacements and velocities the integrand in the above can be written as

$$\frac{\underline{u}^j(r) - \underline{u}^{j-1}(r)}{t_j - t_{j-1}} - \frac{1}{2}\left(\underline{v}^j(r) + \underline{v}^{j-1}(r)\right) = \sum_{i=1}^{2n} e_i \underline{H}_i(r) \tag{6.6.22}$$

where

$$e_i = \left(\frac{c_i^j - c_i^{j-1}}{t_j - t_{j-1}}\right) - \frac{1}{2}\left(b_i^j + b_i^{j-1}\right). \tag{6.6.23}$$

When the index i corresponds to a known boundary value we have $e_i = 0$. This we

consider for all appropriate k for which this must hold which gives

$$\sum_{i=1}^{2n} e_i \int_0^1 (\underline{H}_i \cdot \underline{H}_k)\, r\mathrm{d}r = 0, \quad \forall \text{ appropriate } k. \tag{6.6.24}$$

This is a homogeneous linear system involving a mass matrix and as a mass matrix is positive definite and hence non-singular this implies that

$$e_i = 0, \quad i = 1, \ldots, 2n, \tag{6.6.25}$$

i.e. the displacement and velocity vectors are related by

$$\frac{\underline{u}^j(r) - \underline{u}^{j-1}(r)}{t_j - t_{j-1}} = \frac{1}{2}\left(\underline{v}^j(r) + \underline{v}^{j-1}(r)\right). \tag{6.6.26}$$

Thus in the discrete scheme the time derivative of the displacement matches the velocity at the mid-times in $[t_{j-1}, t_j]$ and in particular the connection between $\underline{u}^j$ and $\underline{v}^j$ is that

$$\underline{v}^j(r) = 2\left(\frac{\underline{u}^j(r) - \underline{u}^{j-1}(r)}{t_j - t_{j-1}}\right) - \underline{v}^{j-1}(r). \tag{6.6.27}$$

The matching of the approximate velocity $\underline{v}$ with the time derivative of the displacement in a weak sense does hence give identifiable points in time when they match in a pointwise sense. This is generalised later in the higher order in time scheme.

Terms that involve only $\underline{\psi}$-functions

Next we consider the "$\underline{\psi}_k$ equations" and to shorten slightly what is written let

$$\tilde{a}(\underline{u}(r,t); \underline{\psi}(r,t)) = h_0 a_1(\underline{u}(r,t); \underline{\psi}(r,t)) - P(t) a_2(\underline{u}(r,t); \underline{\psi}(r,t)) \tag{6.6.28}$$

where $a_1(\cdot;\cdot)$ and $a_2(\cdot,\cdot)$ are as in the quasi-static description. With this shorthand the equations are of the form

$$\begin{aligned}\int_{t_{j-1}}^{t_j} \int_0^1 \tilde{a}(\underline{u}(r,t); \underline{H}_k(r))\; r\mathrm{d}r\mathrm{d}t \\ + \rho h_0 \int_{t_{j-1}}^{t_j} \int_0^1 \left(\frac{\underline{v}^j(r) - \underline{v}^{j-1}(r)}{t_j - t_{j-1}}\right) \cdot \underline{H}_k(r)\; r\mathrm{d}r\mathrm{d}t = 0.\end{aligned} \tag{6.6.29}$$

The integrand in the second integral does not vary with t and the integral can be done exactly but the first integral needs to be approximated and we choose to use the mid-point

rule approximation in time, i.e.

$$\int_{t_{j-1}}^{t_j}\int_0^1 \tilde{a}(\underline{u}(r,t);\underline{H}_k(r))\ r\mathrm{d}r\mathrm{d}t \approx (t_j - t_{j-1})\int_0^1 \tilde{a}(\underline{u}^{j-1/2}(r);\underline{H}_k(r))\ r\mathrm{d}r \tag{6.6.30}$$

where

$$\underline{u}^{j-1/2}(r) := \underline{u}(r,t^{j-1/2}) = \frac{1}{2}\left(\underline{u}^j(r) + \underline{u}^{j-1}(r)\right), \quad \text{with} \quad t^{j-1/2} := (t^j + t^{j-1})/2. \tag{6.6.31}$$

Now by using (6.6.27), the integrand in the second part of (6.6.29) can be expressed in terms of $\underline{u}^j(r)$ to be

$$(t_j - t_{j-1})\,\frac{2\rho h_0}{(t_j - t_{j-1})}\int_0^1 \left(\frac{\underline{u}^j(r) - \underline{u}^{j-1}(r)}{t_j - t_{j-1}} - \underline{v}^{j-1}(r)\right)\cdot \underline{H}_k(r)\ r\mathrm{d}r. \tag{6.6.32}$$

Putting the two parts together gives the equations to solve which determine the parameters $\underline{c}^j$ of the function $\underline{u}^j(r)$ can be written as

$$\begin{aligned}&\int_0^1 \tilde{a}(\underline{u}^{j-1/2}(r);\underline{H}_k(r))\ r\mathrm{d}r\\ &+\frac{2\rho h_0}{(t_j - t_{j-1})}\int_0^1 \left(\frac{\underline{u}^j(r) - \underline{u}^{j-1}(r)}{t_j - t_{j-1}} - \underline{v}^{j-1}(r)\right)\cdot \underline{H}_k(r)\ r\mathrm{d}r = 0\end{aligned} \tag{6.6.33}$$

and these must apply for all appropriate k.

Note The term $(t_j - t_{j-1})$, for both integrals in (6.6.33), is omitted since the right hand side of the system is equal to zero.

This is a system of nonlinear equations which we need to solve for $\underline{c}^j$ by using Newton iteration and we can start the Newton iteration to get $\underline{c}^j$ with the vector $\underline{c}^{j-1}$. The Jacobian matrix of the system has a similar structure to the quasi-static case with an additional mass-matrix contribution arising from the last part of (6.6.33). Once the Newton iteration has converged to give $\underline{c}^j$ we get $\underline{b}^j$ from (6.6.27), i.e. we have

$$\underline{b}^j = 2\left(\frac{\underline{c}^j - \underline{c}^{j-1}}{t_j - t_{j-1}}\right) - \underline{b}^{j-1}. \tag{6.6.34}$$

7. DUAL PROBLEMS FOR AXISYMMETRIC MEMBRANE DEFORMATION

7.1 Introduction

Dual problems that arise in attempts to determine the error in an approximation to a QoI and to have quantities which may help to drive goal orientated adaptive refinement. As we show in section 7.3 the dual problem corresponding to the dynamic case involves solving a problem which is backward in time. If we were to stop at this point then the description would not go much beyond what is given in [28] when the aim in that paper was to predict modelling error. What we might describe as the basic or standard finite element schemes for the forward problem and the backward dual problem, which was used in [28] and which we describe here in sections 6.6 and 7.5, is not sufficiently accurate with respect to the time discretization to do well enough when the context is error estimation. This can be shown from the numerical experiments, which we show in the next chapter for the basic-scheme in time, in subsections (8.3.1) - (8.3.2). From the results, we can conclude that we need to put a lot of effort for the time discretization which is very computationally expensive.

As a consequence of this, we decided to go beyond of what was done in [28] by considering finite element schemes which involve higher order polynomials in time, which is described in the last part of this chapter.

7.2 The dual problem in the quasi-static case

As part of the technique to estimate the error in an approximation to a quantity of interest $J(\underline{u})$, where J denotes a quantity of interest functional, we need to set-up a related dual problem and we consider here the dual problem in the quasi-static case which we write in the following way.

Find $\underline{\psi}$ such that

$$A'_Q(\underline{u};\underline{\alpha},\underline{\psi}) = J'(\underline{u};\underline{\alpha}) \quad \forall \text{ suitable } \underline{\alpha}. \tag{7.2.1}$$

Similar to what is given in (4.6.8) in the non-axisymmetric case the functional J can be of the form

$$J(\underline{u}) = \int_{\Omega^*} \{\text{expressions in } \underline{u} \text{ and } \underline{u}'\}\ rdr. \tag{7.2.2}$$

In the quasi-static case we have

$$A'_Q(\underline{u};\underline{\alpha},\underline{\psi}) = a'(\underline{u};\underline{\alpha},\underline{\psi})_\Omega \tag{7.2.3}$$

where

$$a'(\underline{u};\underline{\alpha},\underline{\psi})_\Omega = h_0 \int_0^1 a'_1(\underline{u};\underline{\alpha},\underline{\psi})\ r\mathrm{d}r - P(t) \int_0^1 a'_2(\underline{u};\underline{\alpha},\underline{\psi})\ r\mathrm{d}r. \tag{7.2.4}$$

The expressions for the Gâteaux derivatives of a_1 and a_2 are a bit shorter than the corresponding ones in section 4.6 and we give these next.

The $a'_1(\underline{u};\underline{\alpha},\underline{\psi})$ expression

We obtain $a'_1(\underline{u};\underline{\alpha},\underline{\psi})$ by taking the Gâteaux derivative of $a_1(\underline{u};\underline{\psi})$ given in (6.3.5) and it helps here to use the following notation for the partial derivatives of the strain energy density W.

$$W_1 = \frac{\partial W}{\partial \lambda_1}, \quad W_2 = \frac{\partial W}{\partial \lambda_2}, \quad W_{11} = \frac{\partial^2 W}{\partial \lambda_1^2}, \quad W_{12} = \frac{\partial^2 W}{\partial \lambda_1 \partial \lambda_2}, \quad W_{22} = \frac{\partial^2 W}{\partial \lambda_2^2}.$$

By using the chain rule we get

$$\begin{aligned} a'_1(\underline{u};\underline{\alpha},\underline{\psi}) &= (W_{11}\lambda'_1(\underline{u};\underline{\alpha}) + W_{12}\lambda'_2(\underline{u};\underline{\alpha}))\lambda'_1(\underline{u};\underline{\psi}) + W_1\lambda''_1(\underline{u};\underline{\alpha},\underline{\psi}) \\ &\quad +(W_{12}\lambda'_1(\underline{u};\underline{\alpha}) + W_{22}\lambda'_2(\underline{u};\underline{\alpha}))\lambda'_2(\underline{u};\underline{\psi}) + W_2\lambda''_2(\underline{u};\underline{\alpha},\underline{\psi}), \\ &= W_{11}\lambda'_1(\underline{u};\underline{\alpha})\lambda'_1(\underline{u};\underline{\psi}) + W_{22}\lambda'_2(\underline{u};\underline{\alpha})\lambda'_2(\underline{u};\underline{\psi}) \\ &\quad +W_{12}(\lambda'_1(\underline{u};\underline{\alpha})\lambda'_2(\underline{u};\underline{\psi}) + \lambda'_1(\underline{u};\underline{\psi})\lambda'_2(\underline{u};\underline{\alpha})) \\ &\quad +W_1\lambda''_1(\underline{u};\underline{\alpha},\underline{\psi}) + W_2\lambda''_2(\underline{u};\underline{\alpha},\underline{\psi}), \end{aligned} \tag{7.2.5}$$

where the Gâteaux derivatives of the stretch ratios are given by

$$\lambda'_1(\underline{u};\underline{\psi}) = \frac{(1+u'_1)\psi'_1 + u'_3\psi'_3}{\lambda_1}, \qquad \lambda'_1(\underline{u};\underline{\alpha}) = \frac{(1+u'_1)\alpha'_1 + u'_3\alpha'_3}{\lambda_1},$$
$$\lambda'_2(\underline{u};\underline{\psi}) = \frac{\psi_1}{r}, \qquad \lambda'_2(\underline{u};\underline{\alpha}) = \frac{\alpha_1}{r}$$

along with

$$\lambda_1''(\underline{u};\underline{\alpha},\underline{\psi}) = \frac{\alpha_1'\psi_1' + \alpha_3'\psi_3' - \lambda_1'(\underline{u};\underline{\alpha})\lambda_1'(\underline{u};\underline{\psi})}{\lambda_1} \quad \text{and} \quad \lambda_2''(\underline{u};\underline{\alpha},\underline{\psi}) = 0. \tag{7.2.6}$$

Thus we have

$$a_1'(\underline{u};\underline{\alpha},\underline{\psi})_\Omega = \int_0^1 (\alpha_1, \alpha_1', \alpha_3') G_1 \begin{pmatrix} \psi_1 \\ \psi_1' \\ \psi_3' \end{pmatrix} \, r\mathrm{d}r, \tag{7.2.7}$$

where

$$G_1 = \begin{pmatrix} \frac{W_{22}}{r^2} & \frac{W_{12}}{r}\frac{(1+u_1')}{\lambda_1} & \frac{W_{12}}{r}\frac{u_3'}{\lambda_1} \\ \frac{W_{12}}{r}\frac{(1+u_1')}{\lambda_1} & W_{11}\frac{(1+u_1')^2}{\lambda_1^2} + W_1\frac{{u_3'}^2}{\lambda_1^3} & \frac{(1+u_1')u_3'}{\lambda_1^2}\left(W_{11} - \frac{W_1}{\lambda_1}\right) \\ \frac{W_{12}}{r}\frac{u_3'}{\lambda_1} & \frac{(1+u_1')u_3'}{\lambda_1^2}\left(W_{11} - \frac{W_1}{\lambda_1}\right) & W_{11}\frac{{u_3'}^2}{\lambda_1^2} + W_1\frac{(1+u_1')^2}{\lambda_1^3} \end{pmatrix}. \tag{7.2.8}$$

The $a_2'(\underline{u};\underline{\alpha},\underline{\psi})$ expression

By using (6.3.7) we get the Gâteaux derivative of $a_2(\underline{u};\underline{\psi})$ as

$$a_2'(\underline{u};\underline{\alpha},\underline{\psi}) = \left(1 + \frac{u_1}{r}\right)(-\alpha_3'\psi_1 + \alpha_1'\psi_3) + \frac{\alpha_1}{r}(-u_3'\psi_1 + (1+u_1')\psi_3). \tag{7.2.9}$$

It is an integral of this that we need and by using integration by parts in the integral involving r we can write this in a form which only involves $\alpha_1, \alpha_1', \alpha_3'$ and ψ_1, ψ_1', ψ_3', as with the previous term, as follows.

$$ra_2'(\underline{u};\underline{\alpha},\underline{\psi}) = ((r+u_1)\alpha_1)'\psi_3 - (r+u_1)\alpha_3'\psi_1 - u_3'\alpha_1\psi_1. \tag{7.2.10}$$

Integrating on $0 \le r < 1$ and using $\alpha_1(1) = 0$ gives

$$\begin{aligned} \int_0^1 a_2'(\underline{u};\underline{\alpha},\underline{\psi})\; r\mathrm{d}r &= -\int_0^1 (r+u_1)\alpha_1\psi_3' + (r+u_1)\alpha_3'\psi_1 + u_3'\alpha_1\psi_1 \,\mathrm{d}r \\ &= -\int_0^1 (\alpha_1, \alpha_1', \alpha_3') G_2 \begin{pmatrix} \psi_1 \\ \psi_1' \\ \psi_3' \end{pmatrix} \, r\mathrm{d}r \end{aligned} \tag{7.2.11}$$

where

$$G_2 = \begin{pmatrix} \frac{u_3'}{r} & 0 & \lambda_2 \\ 0 & 0 & 0 \\ \lambda_2 & 0 & 0 \end{pmatrix}. \tag{7.2.12}$$

Hence in what follows we re-define $a_2'(\underline{u};\underline{\alpha},\underline{\psi})$ as

$$a_2'(\underline{u};\underline{\alpha},\underline{\psi}) = -(\alpha_1,\alpha_1',\alpha_3')G_2 \begin{pmatrix} \psi_1 \\ \psi_1' \\ \psi_3' \end{pmatrix}. \tag{7.2.13}$$

7.3 The dual problem for the dynamic case

The weak form description for the dynamic case is given by (6.4.6)- (6.4.8) and (6.4.9) and to get a corresponding dual problem we need to get the Gâteaux derivative involving changes in $\underline{u}$ and changes in $\underline{v}$. As before we let the changes in $\underline{u}$ be in the direction of $\underline{\alpha}$ and we now let the change in $\underline{v}$ be in the direction of $\underline{\beta}$ with in components $\underline{\alpha} = (\alpha_1,\alpha_3)^T$ and $\underline{\beta} = (\beta_1,\beta_3)^T$. By using these directions we can describe the dual problem as follows.

Find $\underline{\psi}$ and $\underline{\theta}$ such that

$$A'\left(\begin{pmatrix}\underline{u}\\ \underline{v}\end{pmatrix};\begin{pmatrix}\underline{\alpha}\\ \underline{\beta}\end{pmatrix},\begin{pmatrix}\underline{\psi}\\ \underline{\theta}\end{pmatrix}\right) = J'\left(\begin{pmatrix}\underline{u}\\ \underline{v}\end{pmatrix};\begin{pmatrix}\underline{\alpha}\\ \underline{\beta}\end{pmatrix}\right) \quad \forall \text{ suitable } \begin{pmatrix}\underline{\alpha}\\ \underline{\beta}\end{pmatrix}. \tag{7.3.1}$$

Here, $\underline{u}$ and $\underline{v}$ are data used for the dual problems and in the computational scheme they are the finite element approximations $\underline{u}_h$ and $\underline{v}_h$ respectively, or of something derived from these.

For the functional J we assume that is of the following form

$$\begin{aligned} J\left(\begin{pmatrix}\underline{u}\\ \underline{v}\end{pmatrix}\right) &= \int_0^T\int_0^1 r\{\text{expression in } \underline{u},\ \underline{v},\ \underline{u}'\}\,\mathrm{d}r\mathrm{d}t \\ &\quad + \int_0^1 r\{\text{expression in } \underline{u}(\cdot,T),\ \underline{v}(\cdot,T)\}\,\mathrm{d}r. \end{aligned} \tag{7.3.2}$$

This leads to a representation of the Gâteaux derivative of the form

$$J'\left(\begin{pmatrix}\underline{u}\\ \underline{v}\end{pmatrix};\begin{pmatrix}\underline{\alpha}\\ \underline{\beta}\end{pmatrix}\right) = \int_0^T\int_0^1 (\underline{\alpha}\cdot\underline{J}_\alpha + \underline{\beta}\cdot\underline{J}_\beta + \underline{\alpha}'\cdot\underline{J}_{\alpha'})\; r\mathrm{d}r\mathrm{d}t + \int_0^1 (\underline{\alpha}(r,T)\cdot\underline{J}_\alpha(r,T) + \underline{\beta}(r,T)\cdot\underline{J}_\beta(r,T))\; r\mathrm{d}r \quad (7.3.3)$$

where in the above the terms $\underline{J}_\alpha$, $\underline{J}'_\alpha$ and $\underline{J}_\beta$ provide user-defined data to the dual problem and may depend on $\underline{u}$ and $\underline{v}$. Examples of functionals J considered in the next chapter are the average of $u_3(r,T)$ at the final time T near the pole and the average thickness over a time interval $[T-\delta,T]$ for some $\delta>0$ near the pole.

We consider next the expression for A' and it helps here to introduce the notation

$$a'(\underline{u};\underline{\alpha},\underline{\psi})_Q = \int_0^T a'(\underline{u};\underline{\alpha},\underline{\psi})_\Omega\,\mathrm{d}t \quad (7.3.4)$$

$$= h_0\int_0^T\int_0^1 a'_1(\underline{u};\underline{\alpha},\underline{\psi})\; r\mathrm{d}r\mathrm{d}t - \int_0^T P(t)\int_0^1 a'_2(\underline{u};\underline{\alpha},\underline{\psi})\; r\mathrm{d}r\mathrm{d}t. \quad (7.3.5)$$

With this notation we can write the Gâteaux derivative as

$$A'\left(\begin{pmatrix}\underline{u}\\ \underline{v}\end{pmatrix};\begin{pmatrix}\underline{\alpha}\\ \underline{\beta}\end{pmatrix},\begin{pmatrix}\underline{\psi}\\ \underline{\theta}\end{pmatrix}\right) = a'(\underline{u};\underline{\alpha},\underline{\psi})_Q + \rho h_0(\dot{\underline{\beta}},\underline{\psi})_Q \quad (7.3.6)$$

$$+\rho h_0\left((\dot{\underline{\alpha}}-\underline{\beta},\underline{\theta})_Q + (\underline{\alpha}(\cdot,0),\underline{\theta})_\Omega + (\underline{\beta}(\cdot,0),\underline{\psi})_\Omega\right). \quad (7.3.7)$$

As given this involves time derivatives of $\underline{\alpha}$ and $\underline{\beta}$ but there are no time derivatives of these in (7.3.3). To get a matching form we use integration by parts for the integral in t for the $\dot{\underline{\alpha}}$ and $\dot{\underline{\beta}}$ terms and specifically this gives

$$(\dot{\underline{\beta}},\underline{\psi})_Q = (\underline{\beta}(\cdot,T)-\underline{\beta}(\cdot,0),\underline{\psi})_\Omega - (\underline{\beta},\dot{\underline{\psi}})_Q, \quad (7.3.8)$$

$$(\dot{\underline{\alpha}},\underline{\theta})_Q = (\underline{\alpha}(\cdot,T)-\underline{\alpha}(\cdot,0),\underline{\theta})_\Omega - (\underline{\alpha},\dot{\underline{\theta}})_Q. \quad (7.3.9)$$

By using these relations we get

$$A'\left(\begin{pmatrix}\underline{u}\\ \underline{v}\end{pmatrix};\begin{pmatrix}\underline{\alpha}\\ \underline{\beta}\end{pmatrix},\begin{pmatrix}\underline{\psi}\\ \underline{\theta}\end{pmatrix}\right) = a'(\underline{u};\underline{\alpha},\underline{\psi})_Q - \rho h_0(\underline{\beta},\dot{\underline{\psi}})_Q$$

$$-\rho h_0\left((\underline{\alpha},\dot{\underline{\theta}})_Q + (\underline{\beta},\underline{\theta})_Q - (\underline{\alpha}(\cdot,T),\underline{\theta})_\Omega - (\underline{\beta}(\cdot,T),\underline{\psi})_\Omega\right). \quad (7.3.10)$$

7.4 A finite element model for the dual problem in the quasi-static case

As stated a few times already in this thesis, the details in a dual problem can be highly problem dependent. In the case being considered here of the axisymmetric case and a quasi-static deformation much of the detail has already been given in section 7.2 and hence this section is quite short. If the finite element solution $\underline{u}_h$ is obtained by solving (6.5.7) then the finite element dual problem involves solving a problem of the following form.

Find $\underline{\psi}_h \in \bar{V}_h$ such that

$$A'_Q(\tilde{\underline{u}}_h; \underline{\alpha}, \underline{\psi}_h) = J'(\tilde{\underline{u}}_h; \underline{\alpha}) \quad \forall \underline{\alpha} \in \bar{V}_h, \tag{7.4.1}$$

where $\tilde{\underline{u}}_h$ is $\underline{u}_h$ or the average of $\underline{u}_h$ and a better approximation and where the finite element space $\bar{V}_h$ needs to be different to the space V_h used to get $\underline{u}_h$. In our computation our different space $\bar{V}_h$ involves piecewise polynomials of one degree higher than is used for V_h.

7.5 A basic finite element model for the dual problem in the dynamic case

There is a lot more detail to give to describe a scheme to approximately solve the dual problem given in (7.3.1) where $A'(\cdot;\cdot,\cdot)$ is given in (7.3.10) and $J'(\cdot;\cdot)$ is of the form (7.3.3). We do this in this section in a manner similar to [28] which uses degree 1 polynomials in time t in the time intervals and we generalise this in section 7.6 when higher order polynomials in time are used.

In the following we denote the time intervals as $0 = t_0 < t_1 < \cdots < t_N = T$ and typically these are the same time levels as are used to get the finite element displacements and velocity or they are finer discretization with time levels in common. It is important to note that for the dual problem $\underline{u}$ and $\underline{v}$ are given, i.e. they are the data for the problem, and the unknowns are now $\underline{\psi}(r,t)$ and $\underline{\theta}(r,t)$. As in section 6.6 to get the finite element displacement and velocity we let the spatial discretization be fixed in time and we let the basis functions in space be denoted by $\underline{H}_1(r), \ldots, \underline{H}_{2n}(r)$. Then we define the following approximations for each time level in the form

$$\underline{\psi}^j(r) := \sum_{i=1}^{2n} c_i^j \underline{H}_i(r), \quad \underline{\theta}^j(r) := \sum_{i=1}^{2n} b_i^j \underline{H}_i(r). \tag{7.5.1}$$

In our application here $\underline{H}_1, \ldots, \underline{H}_{2n}$ are usually higher degree piecewise polynomials than was used to get $\underline{u}$ and $\underline{v}$. As before, between time levels, i.e. $t_{j-1} < t < t_j$ we assume that the dependence on t is of the form

$$\underline{\psi}(r,t) = \left(\frac{t_j - t}{t_j - t_{j-1}}\right)\underline{\psi}^{j-1}(r) + \left(\frac{t - t_{j-1}}{t_j - t_{j-1}}\right)\underline{\psi}^{j}(r), \tag{7.5.2}$$

$$\underline{\theta}(r,t) = \left(\frac{t_j - t}{t_j - t_{j-1}}\right)\underline{\theta}^{j-1}(r) + \left(\frac{t - t_{j-1}}{t_j - t_{j-1}}\right)\underline{\theta}^{j}(r). \tag{7.5.3}$$

This is the degree 1 polynomial in t behaviour which we generalise in section 7.6.

One of the main differences between solving the dual problem compared to solving the problem to get $\underline{u}$ and $\underline{v}$ is that we have to solve backward in time. Thus, we start with the final time conditions at $t = T$ which determine $\underline{c}^T$ and $\underline{b}^T$. For this case when the solution is known at time t_j, and thus we know $\underline{c}^j$ and $\underline{b}^j$, the unknowns are $\underline{c}^{j-1}$ and $\underline{b}^{j-1}$. The discrete version of (7.3.1) to determine $\underline{c}^{j-1}$ and $\underline{b}^{j-1}$ depends on which test vectors $\underline{\alpha}$ and $\underline{\beta}$ which are used for each case. As in section 6.6, in the case of $\underline{\alpha}$ we take the functions

$$\underline{\alpha}_k(r,t) = \begin{cases} \underline{H}_k(r), & t_{j-1} < t < t_j, \\ 0, & \text{otherwise}, \end{cases} \tag{7.5.4}$$

and for $\underline{\beta}$ we take

$$\underline{\beta}_k(r,t) = \begin{cases} \underline{H}_k(r), & t_{j-1} < t < t_j, \\ 0, & \text{otherwise}. \end{cases} \tag{7.5.5}$$

Note that these functions do not vary with t on the time interval. In both cases we take all the values of k with $k = 1, \ldots, 2n$ except those corresponding to known boundary values at $r = 0$ or at $r = 1$. We separately have to satisfy

$$A'\left(\begin{pmatrix}\underline{u}\\ \underline{v}\end{pmatrix}; \begin{pmatrix}\underline{\alpha}_k\\ \underline{0}\end{pmatrix}, \begin{pmatrix}\underline{\psi}\\ \underline{\theta}\end{pmatrix}\right) = J'\left(\begin{pmatrix}\underline{u}\\ \underline{v}\end{pmatrix}; \begin{pmatrix}\underline{\alpha}_k\\ \underline{0}\end{pmatrix}\right) \quad \forall \text{ appropriate } k, \tag{7.5.6}$$

$$A'\left(\begin{pmatrix}\underline{u}\\ \underline{v}\end{pmatrix}; \begin{pmatrix}\underline{0}\\ \underline{\beta}_k\end{pmatrix}, \begin{pmatrix}\underline{\psi}\\ \underline{\theta}\end{pmatrix}\right) = J'\left(\begin{pmatrix}\underline{u}\\ \underline{v}\end{pmatrix}; \begin{pmatrix}\underline{0}\\ \underline{\beta}_k\end{pmatrix}\right) \quad \forall \text{ appropriate } k. \tag{7.5.7}$$

The different parts when solving the dual problem

The final time conditions at $t = T$

From the terms involving $\underline{\alpha}(\cdot, T)$ and $\underline{\beta}(\cdot, T)$ we have the following two systems that we need to solve for $\underline{\theta}(\cdot, T)$ and $\underline{\psi}(\cdot, T)$ respectively.

To obtain $\underline{\theta}(\cdot, T)$ we do the following. We consider the test vector with $\underline{\beta}_k = \underline{0}$ and $\underline{\alpha}_k \neq \underline{0}$ where we have the following system:
Find $\underline{\theta}(\cdot, T)$ such that

$$\rho h_0(\underline{\alpha}_k(\cdot, T), \underline{\theta}(\cdot, T))_\Omega = J'\left(\begin{pmatrix}\underline{u}\\ \underline{v}\end{pmatrix}; \begin{pmatrix}\underline{\alpha}_k\\ \underline{0}\end{pmatrix}\right) \quad \forall \text{ appropriate } k. \tag{7.5.8}$$

In full this gives

$$\rho h_0(\underline{\alpha}(\cdot, T), \underline{\theta}(\cdot, T))_\Omega = \int_0^1 (\underline{\alpha}_k(r, T) \cdot \underline{J}_\alpha(r, T))\, r\mathrm{d}r. \tag{7.5.9}$$

To obtain $\underline{\psi}(\cdot, T)$ we do the following.
We consider the test vector with $\underline{\beta}_k \neq \underline{0}$ and $\underline{\alpha}_k = \underline{0}$ where we have the following system:
Find $\underline{\psi}(\cdot, T)$ such that

$$\rho h_0(\underline{\beta}(\cdot, T), \underline{\psi}(\cdot, T))_\Omega = J'\left(\begin{pmatrix}\underline{u}\\ \underline{v}\end{pmatrix}; \begin{pmatrix}\underline{0}\\ \underline{\beta}_k\end{pmatrix}\right) \quad \forall \text{ appropriate } \underline{\beta}. \tag{7.5.10}$$

Similar to the previous case this gives

$$\rho h_0(\underline{\beta}(\cdot, T), \underline{\psi}(\cdot, T))_\Omega = \int_0^1 \underline{\beta}_k(r, T) \cdot \underline{J}_\beta(r, T)\, r\mathrm{d}r. \tag{7.5.11}$$

Solving in the time interval $t_{j-1} < t < t_j$

The equations involving the $\underline{\beta}$ test functions

From the terms involving $\underline{\beta}(\cdot, t_j)$ test functions, which are non-zero only on $t_{j-1} \le t \le t_j$, we get the following equations which involve $\underline{\psi}(\cdot, t_j)$ and $\underline{\theta}(\cdot, t_j)$ for all the time intervals.

$$-\rho h_0(\dot{\underline{\psi}} + \underline{\theta}, \underline{\beta}_k)_Q = J'\left(\begin{pmatrix}\underline{u}\\ \underline{v}\end{pmatrix}; \begin{pmatrix}\underline{0}\\ \underline{\beta}_k\end{pmatrix}\right) = \int_{t_{j-1}}^{t_j}\int_0^1 \underline{\beta}_k \cdot \underline{J}_\beta r dr dt, \ \forall \text{ appropriate } k \ . \tag{7.5.12}$$

These "$\underline{\beta}_k$ equations" give a connection between the vectors $\underline{b}^j$ and $\underline{c}^j$ similar to the relation between the displacement and the velocity. In full these are

$$-\rho h_0 \int_{t_{j-1}}^{t_i}\int_0^1 \left(\dot{\underline{\psi}}(r,t) + \underline{\theta}(r,t)\right) \cdot \underline{\beta}_k(r)\ r\mathrm{d}r dt = \int_{t_{j-1}}^{t_i}\int_0^1 \underline{\beta}_k \cdot \underline{J}_\beta r dr dt \quad \forall \text{ appropriate } k. \tag{7.5.13}$$

For $t_{j-1} < t < t_j$ the term $\dot{\underline{\psi}}$ does not vary with t and the $\underline{\theta}(r,t)$ varies with t as described in (7.5.3) and we can do both of the integrals in time exactly to give

$$\begin{aligned} & -\rho h_0 (t_j - t_{j-1}) \int_0^1 \left(\frac{\underline{\psi}^j(r) - \underline{\psi}^{j-1}(r)}{t_j - t_{j-1}} + \frac{1}{2}\left(\underline{\theta}^j(r) + \underline{\theta}^{j-1}(r)\right)\right) \cdot \underline{H}_k(r)\ r\mathrm{d}r \\ = \ & (t_j - t_{j-1}) \int_0^1 \underline{H}_k \cdot \underline{J}_\beta\, r dr. \end{aligned}$$

In the simplest case when $\underline{J}_\beta = \underline{0}$ and the reasoning used which gave (6.6.26) gives here that

$$\frac{\underline{\psi}^j(r) - \underline{\psi}^{j-1}(r)}{t_j - t_{j-1}} + \frac{1}{2}\left(\underline{\theta}^j(r) + \underline{\theta}^{j-1}(r)\right) = \underline{0}. \tag{7.5.14}$$

In the examples considered in this thesis we only consider functionals $J(\cdot)$ such that $\underline{J}_\beta = \underline{0}$. If $\underline{J}_\beta \neq \underline{0}$ then we construct a function

$$\underline{\gamma}^j(r) = \sum_{i=1}^{2n} e_i \underline{H}_i(r) \tag{7.5.15}$$

such that

$$\int_0^1 \underline{\gamma}^j(r) \cdot \underline{H}_k r\, dr = \int_0^1 \underline{\beta}_k \cdot \underline{J}_\beta r\, dr \quad \forall \text{appropriate } k. \tag{7.5.16}$$

To satisfy (7.5.13) requires that (7.5.14) is replaced by

$$\frac{\underline{\psi}^j(r) - \underline{\psi}^{j-1}(r)}{t_j - t_{j-1}} + \frac{1}{2}\left(\underline{\theta}^j(r) + \underline{\theta}^{j-1}(r)\right) = \underline{\gamma}^j(r) \tag{7.5.17}$$

and in terms of components

$$\left(\frac{c_i^j - c_i^{j-1}}{t_j - t_{j-1}}\right) + \frac{1}{2}\left(b_i^j + b_i^{j-1}\right) = e_i. \tag{7.5.18}$$

A consequence of this is that when we are solving on $[t_{j-1}, t_j]$ we already know $\underline{\psi}^j$ and $\underline{\theta}^j$ and we are trying to determine $\underline{\psi}^{j-1}$ and $\underline{\theta}^{j-1}$ and we have

$$\underline{\theta}^{j-1}(r) = -2\left(\frac{\underline{\psi}^j(r) - \underline{\psi}^{j-1}(r)}{t_j - t_{j-1}}\right) - \underline{\theta}^j(r) + 2\underline{\gamma}^j(r). \tag{7.5.19}$$

We use this in a moment to get equations which are just in terms of $\underline{\psi}^{j-1}$.

The equations involving the $\underline{\alpha}$ test functions

Collecting the terms that involve $\underline{\alpha}(\cdot, t_j)$ test functions, which are only non zero on $t_{j-1} \leq t \leq t_j$ we get the following

$$a'(\underline{u}; \underline{\alpha}_k, \underline{\psi})_Q - \rho h_0(\dot{\underline{\theta}}, \underline{\alpha}_k)_Q = J'\left(\begin{pmatrix}\underline{u}\\ \underline{v}\end{pmatrix}; \begin{pmatrix}\underline{\alpha}_k\\ \underline{0}\end{pmatrix}\right) = \int_{t_{j-1}}^{t_j}\int_0^1 (\underline{\alpha}\cdot\underline{J}_\alpha + \underline{\alpha}_k'\cdot\underline{J}_{\alpha'})r\,drdt. \tag{7.5.20}$$

In full these are as follows.

$$\begin{aligned}&\int_{t_{j-1}}^{t_j}\int_0^1 a'(\underline{u}(r,t); \underline{\alpha}_k(r), \underline{\psi}(r,t))rdrdt - \rho h_0\int_{t_{j-1}}^{t_j}\int_0^1 \dot{\underline{\theta}}(r,t)\cdot\underline{\alpha}_k(r)rdrdt\\ &= \int_{t_{j-1}}^{t_j}\int_0^1 (\underline{\alpha}_k\cdot\underline{J}_{\underline{\alpha}} + \underline{\alpha}_k'\cdot\underline{J}_{\underline{\alpha}'})rdrdt.\end{aligned} \tag{7.5.21}$$

By using (7.5.19) we have

$$\underline{\theta}^j - \underline{\theta}^{j-1} = 2\underline{\theta}^j + 2\left(\frac{\underline{\psi}^j - \underline{\psi}^{j-1}}{t_j - t_{j-1}}\right) - 2\underline{\gamma}^j \tag{7.5.22}$$

and thus

$$-\rho h_0(\dot{\underline{\theta}}, \underline{\alpha}_k) = 2\rho h_0 \int_0^1 \left(-\left(\frac{\underline{\psi}^j - \underline{\psi}^{j-1}}{t_j - t_{j-1}} \right) - \underline{\theta}^j + \underline{\gamma}^j \right) \cdot \underline{\alpha}_k r \, dr. \tag{7.5.23}$$

Now for the term involving $a'(\cdot;\cdot,\cdot)$ we need to approximate the integrand in $a'(\underline{u};\underline{\alpha}_k,\underline{\psi})_Q$ and in our scheme we use the mid-point rule so we have the following approximation

$$a'(\underline{u};\underline{\alpha}_k,\underline{\psi})_Q = \int_{t_{j-1}}^t \int_0^1 a'(\underline{u}(r,t);\underline{\alpha}_k(r),\underline{\psi}(r,t_j))drdt \tag{7.5.24}$$

$$\approx (t_j - t_{j-1}) \int_0^1 a'(\underline{u}^{j-1/2}(r);\underline{\alpha}_k(r),\underline{\psi}^{j-1/2}(r))r \, dr, \tag{7.5.25}$$

where as before $t^{j-1/2} := (t^j + t^{j-1})/2$ and

$$\begin{aligned} \underline{u}^{j-1/2}(r) &:= \underline{u}(r,t^{j-1/2}) = \frac{1}{2}(\underline{u}^j(r) + \underline{u}^{j-1}(r)), \\ \underline{\psi}^{j-1/2}(r) &:= \underline{\psi}(r,t^{j-1/2}) = \frac{1}{2}(\underline{\psi}^j(r) + \underline{\psi}^{j-1}(r)) \end{aligned}$$

where $\underline{u}$ is the data used in the dual problem and it is typically $\underline{u}_h$ or something derived from $\underline{u}_h$.

Solving the system to $\underline{\psi}^{j-1}$ and getting $\underline{\theta}^{j-1}$

Now, by using (7.5.23) and (7.5.25) we get the following system which involves only $\underline{\alpha}_k$-test functions

$$\begin{aligned} &(t_j - t_{j-1}) \int_0^1 a' \left(\underline{u}^{j-1/2}(r);\underline{\alpha}_k(r),\underline{\psi}^{j-1/2}(r) \right) r \, dr \\ &\quad +2\rho h_0 \int_0^1 \left(-\left(\frac{\underline{\psi}^j - \underline{\psi}^{j-1}}{t_j - t_{j-1}} \right) - \underline{\theta}^j + \underline{\gamma}^j \right) \cdot \underline{\alpha}_k r \, dr. \\ &\quad = \int_{t_{j-1}}^{t_j} \int_0^1 (\underline{\alpha}_k \cdot \underline{J}_\alpha + \underline{\alpha}'_k \cdot \underline{J}_{\alpha'}) r \, drdt. \end{aligned} \tag{7.5.26}$$

Collecting the terms which involve $\underline{\psi}^{j-1}$ on the left hand side and putting all other terms on the right hand side gives

$$\frac{t_j - t_{j-1}}{2} \int_0^1 a'\left(\underline{u}^{j-1/2}(r); \underline{\alpha}_k, \underline{\psi}^{j-1}(r)\right) r\, dr + \frac{2\rho h_0}{t_j - t_{j-1}} \int_0^1 \underline{\psi}^{j-1} \cdot \underline{\alpha}_k r\, dr \tag{7.5.27}$$

$$= -\frac{t_j - t_{j-1}}{2} \int_0^1 a'\left(\underline{u}^{j-1/2}(r); \underline{\alpha}_k, \underline{\psi}^{j}(r)\right) r\, dr + 2\rho h_0 \int_0^1 \left(\frac{\underline{\psi}^j}{t_j - t_{j-1}} + \underline{\theta}^j - \underline{\gamma}^j\right) \cdot \underline{\alpha}_k r\, dr$$

$$+ \int_{t_{j-1}}^{t_j} \int_0^1 (\underline{\alpha}_k \cdot \underline{J}_\alpha + \underline{\alpha}'_k \cdot \underline{J}_{\alpha'}) r\, dr dt. \tag{7.5.28}$$

This gives a linear system the coefficients of $\underline{\psi}^{j-1}$ and once these are obtained then we substitute back to (7.5.19) to get $\underline{\theta}^{j-1}$.

Remark

The above approach for solving the linear dual problem for $\underline{\psi}$ and $\underline{\theta}$ has a similar structure to the problem of solving the equations of motion to determine the displacement $\underline{u}$ and the velocity $\underline{v}$, with the main difference being that for the dual problem we have final time conditions ($t = T$) instead of initial conditions ($t = 0$). The relationship between $\underline{\psi}$ and $\underline{\theta}$ (7.5.19) is similar to the relation between $\dot{\underline{u}}$ and $\underline{v}$, see (6.6.27), where the condition $\dot{\underline{u}} = \underline{v}$ is imposed weakly.

7.6 A higher order finite element scheme in time for the dynamic case

The reason for setting up and solving a dual problem is to generate functions $\underline{\psi}$ and $\underline{\theta}$ which we will hopefully lead to a sufficiently good estimate of the error in an approximation to a quantity of interest of the form

$$J\begin{pmatrix}\underline{u}\\ \underline{v}\end{pmatrix} - J\begin{pmatrix}\underline{u}_h\\ \underline{v}_h\end{pmatrix} \approx F\begin{pmatrix}\underline{\psi}\\ \underline{\theta}\end{pmatrix} - A\left(\begin{pmatrix}\underline{u}_h\\ \underline{v}_h\end{pmatrix}; \begin{pmatrix}\underline{\psi}\\ \underline{\theta}\end{pmatrix}\right). \tag{7.6.1}$$

We would also like the approximations to be such that the error is small. Unfortunately, with the basic schemes described in sections 6.6 and 7.5 the error is typically dominated by the error due to how we approximate in the time t and to get good accuracy we either have to use a large number of time steps or we need to consider a higher order scheme in time, In this section we consider a higher order scheme in time which generalises what was done in section 6.6.

When we approximate $\underline{u}$ and $\underline{v}$ our generalisation of the scheme in section 6.6 is to

now use polynomials of degree n in time in an interval $[t_{j-1}, t_j]$ and which we use the usual piecewise polynomial finite element functions for the dependence on the radial dimension $r \in [0,1]$ as before. The time scheme which was used before corresponds to the case $n = 1$ and the scheme described here is hence a generalisation.

As before, $\underline{u}(r,t) = (u_1(r,t), u_3(r,t))^T$ denotes the approximate displacement and $\underline{v}(r,t) = (v_1(r,t), v_3(r,t))^T$ is the approximate velocity and in the scheme we impose the condition between $\underline{v}$ and $\dot{\underline{u}}$ weakly which, as we show, requires that the difference $\underline{v}(r,t) - \dot{\underline{u}}(r,t)$ can be described using a Legendre polynomial of degree n with respect to the interval $[t_{j-1}, t_j]$ for the time dependence. An outcome of this is that we can express the equations that we have to solve in terms of only the parameters connected with $\underline{u}(r,t)$. A similar set-up can also be done for the associated dual problem and brief details of this case are given in section 7.7.

7.6.1 Representing $\underline{u}(r,t)$ and $\underline{v}(r,t)$ and the basis functions for the general case

For the description given here we consider the stage when we have the solution at time t_{j-1} and we seek the solution in $t_{j-1} \le t \le t_j$. As notation we let

$$\underline{u}^{j-1}(r) := \underline{u}(r, t_{j-1}), \qquad \underline{v}^{j-1}(r) := \underline{v}(r, t_{j-1}), \tag{7.6.2}$$

$$\underline{u}^{j}(r) := \underline{u}(r, t_{j}), \qquad \underline{v}^{j}(r) := \underline{v}(r, t_{j}). \tag{7.6.3}$$

To describe the time dependence of $\underline{u}(r,t)$ for $t_{j-1} \le t \le t_j$ when $n > 1$ needs additional functions which we give in a moment.

The basis functions that we use for our approximation $\underline{u}(r,t)$ involve the product of two functions, a function of r and a function of t and we start by defining these functions of one variable and what is given here corresponds to the basis functions described in section 2.4.2. In the case of the t dependence and degree $n \ge 1$ polynomials we take

$$\eta_0^n(t) := \frac{t_j - t}{t_j - t_{j-1}}, \quad \eta_n^n(t) := \frac{t - t_{j-1}}{t_j - t_{j-1}}, \qquad \eta_i^n(t) = P_{i+1}(t) - P_{i-1}(t), \quad i = 1, 2, \ldots, n-1, \tag{7.6.4}$$

where here

$$P_i(t) = \hat{P}_i\left(\frac{2t - (t_{j-1} + t_j)}{t_j - t_{j-1}}\right),$$

where $\hat{P}_i$ is the usual Legendre polynomial of degree i on $[-1,1]$. In particular this choice means that the functions $\eta_1^n(t), \ldots, \eta_{n-1}^n(t)$ are 0 when we evaluate at the times t_{j-1} and t_j. In the case of the r dependence and degree $p \ge 1$ polynomials the functions that we

use on an element $[r_s, r_{s+1}]$ are similarly given as

$$\overline{\eta}_0^p(r) := \frac{r_{s+1}-r}{r_{s+1}-r_s}, \quad \overline{\eta}_p^p(r) := \frac{r-r_s}{r_{s+1}-r_s}, \qquad \overline{\eta}_i^p(r) = \tilde{P}_{i+1}(r) - \tilde{P}_{i-1}(r), \quad i = 1, 2, \ldots, p-1, \tag{7.6.5}$$

where here

$$\tilde{P}_i(r) = \hat{P}_i\left(\frac{2r-(r_s+r_{s+1})}{r_{s+1}-r_s}\right)$$

where, as above, $\hat{P}_i$ is the usual Legendre polynomial of degree i on $[-1,1]$. In the notation in both cases the subscript indicates which function is being considered and the superscript indicates the highest degree of all the functions in the basis. The use of the superscript here is because we will also need a basis for polynomials of degree $n-1$ in t when the test functions are introduced. To cover all cases we hence also need the space of degree 0 functions to cover the case when $n = 1$ and in this case we just have

$$\eta_0^0(t) := 1, \quad t_{j-1} < t < t_j. \tag{7.6.6}$$

When we are considering $r_s \le r \le r_{s+1}$ and $t_{j-1} \le t \le t_j$ the solution at time t_{j-1} is already known and the functions $\underline{u}^{j-1}(r)$ and $\underline{v}^{j-1}(r)$ are such that

$$\underline{u}^{j-1}(r), \quad \underline{v}^{j-1}(r) \in \text{span}\left\{\begin{pmatrix}\overline{\eta}_0^p(r)\\0\end{pmatrix}, \begin{pmatrix}0\\\overline{\eta}_0^p(r)\end{pmatrix}, \ldots, \begin{pmatrix}\overline{\eta}_p^p(r)\\0\end{pmatrix}, \begin{pmatrix}0\\\overline{\eta}_p^p(r)\end{pmatrix}\right\}. \tag{7.6.7}$$

We are now ready to describe $\underline{u}(r,t)$ for any r, t on the space-time element as

$$\underline{u}(r,t) = \underline{u}^{j-1}(r)\eta_0^n(t) + \sum_{i=0}^{p}\overline{\eta}_i^p(r)\sum_{l=0}^{n-1}\begin{pmatrix}c_{2(in+l)}\\c_{2(in+l)+1}\end{pmatrix}\eta_{l+1}^n(t). \tag{7.6.8}$$

There are $2(p+1)n$ coefficients labelled as $c_0, c_1, \ldots, c_{2(p+1)n-1}$ which are not known when we start the calculation in the time interval $t_{j-1} \le t \le t_j$. The basis functions on the element are

$$\begin{pmatrix}\overline{\eta}_i^p(r)\eta_{l+1}^n(t)\\0\end{pmatrix}, \quad \begin{pmatrix}0\\\overline{\eta}_i^p(r)\eta_{l+1}^n(t)\end{pmatrix}, \qquad l = 0, \ldots, n-1, \quad i = 0, \ldots, p.$$

These are also the basis for the unknown part of $\underline{v}(r,t)$ on the element at the same stage of the computations although, as we describe, we will express $\underline{v}$ in terms of $\dot{\underline{u}}$ and other terms.

As we explain further in the next section, we have the same number of non-zero test

vectors on the element and these are similarly given by

$$\begin{pmatrix} \overline{\eta}_i^p(r)\eta_l^{n-1}(t) \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ \overline{\eta}_i^p(r)\eta_l^{n-1}(t) \end{pmatrix}, \qquad l = 0, \ldots, n-1, \quad i = 0, \ldots, p.$$

This corresponds to what is done in section 6.6 when it is degree $n = 1$ in time for $\underline{u}(r,t)$ and $\underline{v}(r,t)$ and it is degree 0 in time for the test vectors.

To end this section on representing $\underline{u}(r,t)$ we consider now the number of parameters involved when we consider the complete space region $0 \leq r \leq 1$ when we have ne elements

$$0 = r_0 < r_1 < r_2 < \cdots < r_{ne} = 1.$$

For functions such as $\underline{u}^{j-1}(r)$ and $\underline{v}^{j-1}(r)$ we need 2 scalar parameters at each node r_s and we need $2(p-1)$ interior parameters in each interval (r_s, r_{s+1}). In total we have

$$2(ne+1) + 2(p-1)ne = 2(p(ne)+1). \tag{7.6.9}$$

(This number was given when the quasi-static case was described in section 6.5.) For this description with index values starting at 0, we store these so that the $2(p+1)$ parameters on element $[r_s, r_{s+1}]$ start as index position

$$s(2p), \quad s = 0, 1, \ldots, ne-1.$$

For functions such as $\underline{u}(r,t)$ we need $\underline{u}^{j-1}(r)$ to already be available, we need a further $2n$ scalar parameters at each point r_s and we need $2n(p-1)$ interior parameters in each interval (r_s, r_{s+1}). In total this is n times what is given in (7.6.9), i.e. the storage of these terms has size

$$2n(p(ne)+1).$$

We store these in a column vector so that the $2n(p+1)$ parameters on element $[r_s, r_{s+1}]$ start at index position

$$s(2np), \quad s = 0, 1, \ldots, ne-1.$$

In the next section we discuss the equations that have to be solved and just note here that once a solution is obtained we extract $2(p(ne)+1)$ of the $2n(p(ne)+1)$ parameters to get the parameters describing $\underline{u}^j(r)$ which is needed in the next time interval $[t_j, t_{j+1}]$. For the next time interval we also need the parameters describing $\underline{v}^j(r)$ but before we can discuss this we need the connection between $\underline{\dot{u}}$ and $\underline{v}$ in the approximate scheme and this is the topic of the next section.

7.6.2 The equations to solve for the dynamic problem

In the previous section we defined the form of the approximations $\underline{u}(r,t)$ and $\underline{v}(r,t)$ on the region $0 \le r \le 1$, $t_{j-1} \le t \le t_j$ with the t dependence being a polynomial of degree less than or equal to n. In the description we also noted that $\underline{u}(r,t_{j-1}) = \underline{u}^{j-1}(r)$ and $\underline{v}(r,t_{j-1}) = \underline{v}^{j-1}(r)$ are already known at this stage of the calculation and that there are $2n(p(ne)+1)$ parameters to determine before we take account of boundary conditions at $r=0$ and at $r=1$. We need a matching number of test vectors and these can be defined element-by-element to involve on the element $[r_s, r_{s+1}]$ the span of the functions

$$\begin{pmatrix} \overline{\eta}_i^p(r)\eta_l^{n-1}(t) \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ \overline{\eta}_i^p(r)\eta_l^{n-1}(t) \end{pmatrix}, \qquad l = 0,1,\ldots,n-1, \quad i = 0,1,\ldots,p. \quad (7.6.10)$$

The functions corresponding to $i=0$ when $s>0$ and the functions corresponding to $i=p$ when $s+1<ne$ are part of the description of a function also non-zero on a neighbouring element. In all cases the functions corresponding to $1 \le i \le p-1$ are only non-zero on (r_s, r_{s+1}). A key point here is that the time dependence of the test functions involves all polynomials of one degree lower than that which is used to describe $\underline{u}(r,t)$ and $\underline{v}(r,t)$. We will refer to this property in a moment.

The equations to determine $\underline{u}(r,t)$ and $\underline{v}(r,t)$ in $0 \le r \le 1$, $t_{j-1} \le t \le t_j$ are

$$\int_{t_{j-1}}^{t_j} \int_0^1 (\underline{v}(r,t) - \dot{\underline{u}}(r,t)) \cdot \underline{q}(r,t) r \,\mathrm{d}r\mathrm{d}t = 0, \quad (7.6.11)$$

$$\int_{t_{j-1}}^{t_j} \int_0^1 \left(\tilde{a}(\underline{u}(r,t);\ \underline{q}(r,t)) + \rho h_0 \dot{\underline{v}}(r,t) \cdot \underline{q}(r,t)\right) r \,\mathrm{d}r\mathrm{d}t = 0, \quad (7.6.12)$$

where in each case it is for all $\underline{q}$ in the test space as indicated (7.6.10). As it was described in section 6.6 the first equation is concerned with imposing the condition between $\underline{v}(r,t)$ and $\dot{\underline{u}}(r,t)$ weakly. In the next section we show how this gives an explicit expression between $\dot{\underline{u}}$ and $\underline{v}$ which we can differentiate to give an expression for $\dot{\underline{v}}$ in terms of $\ddot{\underline{u}}$ and other quantities and this enables us to set things up so that we just solve for $\underline{u}$ at any given stage.

7.6.3 *Expressing $\underline{v}$ in terms of $\underline{\dot{u}}$*

Equation (7.6.11) is concerned with imposing the condition between $\underline{v}(r,t)$ and $\underline{\dot{u}}(r,t)$ weakly. This can be solved explicitly by noting that for any $r \in [0,1]$ we satisfy

$$\int_{t_{j-1}}^{t_j} (\underline{v}(r,t) - \underline{\dot{u}}(r,t)) \cdot \underline{q}(r,t)\,\mathrm{d}t = 0$$

for all $\underline{q}$ in the test space with

$$\underline{v}(r,t) - \underline{\dot{u}}(r,t) = \underline{\gamma}(r)P_n(t), \tag{7.6.13}$$

where $P_n(t)$ is the Legendre polynomial of degree n on $[t_{j-1}, t_j]$. This follows because each of the components of the vector $\underline{v}(r,t) - \underline{\dot{u}}(r,t)$ is a polynomial in t of degree less than or equal to n and it is orthogonal to all polynomials of degree $n-1$. We consider next $\underline{\gamma}(r)$. In an iteration to solve the nonlinear equations we have a candidate for the parameters $c_0, c_1, \ldots, c_{2(p+1)n-1}$ on an element and, to repeat, we already know the parameters for $\underline{u}^{j-1}(r)$ and $\underline{v}^{j-1}(r)$. Thus by considering (7.6.13) when $t = t_{j-1}$ we have

$$\underline{v}^{j-1}(r) - \underline{\dot{u}}(r,t_{j-1}) = \underline{\gamma}(r)P_n(t_{j-1}).$$

Now a property of the Legendre polynomial is that $P_n(t_{j-1}) = (-1)^n$ and hence

$$\begin{aligned}
\underline{\gamma}(r) &= (-1)^n \left(\underline{v}^{j-1}(r) - \underline{\dot{u}}(r,t_{j-1})\right) = (-1)^{n+1}\left(\underline{\dot{u}}(r,t_{j-1}) - \underline{v}^{j-1}(r)\right) && (7.6.14)\\
&= (-1)^{n+1}\Bigg(\underline{u}^{j-1}(r)\dot{\eta}_0(t_{j-1}) - \underline{v}^{j-1}(r) \\
&\qquad + \sum_{i=0}^{p} \overline{\eta}_i^p(r) \sum_{l=0}^{n-1} \begin{pmatrix} c_{2(in+l)} \\ c_{2(in+l)+1} \end{pmatrix} \dot{\eta}_{l+1}^n(t_{j-1}).\Bigg) && (7.6.15)
\end{aligned}$$

From this we can get the parameters in a representation of $\underline{\gamma}(r)$ from the space described in (7.6.7) if needed although as we give below we give explicit expressions for the quantities we need to compute which are $\underline{\dot{v}}(r,t)$ during the iteration and $\underline{v}^j(r)$ after the iteration.

7.6.4 *The form of $\underline{\dot{v}}(r,t)$ on an element*

Equation (7.6.12) involves $\underline{\dot{v}}(r,t)$ and this can be written in the form

$$\underline{\dot{v}}(r,t) = \underline{\ddot{u}}(r,t) + \underline{\gamma}(r)\dot{P}_n(t). \tag{7.6.16}$$

On an element we can express this as

$$\begin{aligned}
\underline{\dot{v}}(r,t) &= \underline{u}^{j-1}(r)\ddot{\eta}_0^n(t) + \sum_{i=0}^{p}\overline{\eta}_i^p(r)\sum_{l=0}^{n-1}\begin{pmatrix} c_{2(in+l)} \\ c_{2(in+l)+1}\end{pmatrix}\ddot{\eta}_{l+1}^n(t) \\
&\quad +(-1)^{n+1}\Bigg(\underline{u}^{j-1}(r)\dot{\eta}_0^n(t_{j-1}) - \underline{v}^{j-1}(r) \\
&\quad\quad + \sum_{i=0}^{p}\overline{\eta}_i^p(r)\sum_{l=0}^{n-1}\begin{pmatrix} c_{2(in+l)} \\ c_{2(in+l)+1}\end{pmatrix}\dot{\eta}_{l+1}^n(t_{j-1})\Bigg)\dot{P}_n(t) \\
&= \sum_{i=0}^{p}\overline{\eta}_i^p(r)\sum_{l=0}^{n-2}\begin{pmatrix} c_{2(in+l)} \\ c_{2(in+l)+1}\end{pmatrix}\ddot{\eta}_{l+1}^n(t) \\
&\quad +(-1)^{n+1}\Bigg(\underline{u}^{j-1}(r)\dot{\eta}_0^n(t_{j-1}) - \underline{v}^{j-1}(r) \\
&\quad\quad + \sum_{i=0}^{p}\overline{\eta}_i^p(r)\sum_{l=0}^{n-1}\begin{pmatrix} c_{2(in+l)} \\ c_{2(in+l)+1}\end{pmatrix}\dot{\eta}_{l+1}^n(t_{j-1})\Bigg)\dot{P}_n(t)
\end{aligned}$$

where the final version is because the second time derivatives of the first and last basis functions are 0. This shows explicitly the dependence of $\underline{\dot{v}}(r,t)$ on $c_0, c_1, \ldots, c_{2(p+1)n-1}$ and we refer to this in the next section when we give the element Jacobian matrix needed as part of the things to compute in a Newton iteration.

7.6.5 The element residual vector and the element Jacobian matrix

The detail so far has been about evaluating the functions involved that appear in (7.6.11) and we now considering the integrals that appear in (7.6.12). On an element $[r_s, r_{s+1}]$ we need to compute a quadrature approximation to

$$\int_{t_{j-1}}^{t_j}\int_{r_s}^{r_{s+1}}\left(\tilde{a}(\underline{u}(r,t);\ \underline{q}(r,t)) + \rho h_0\underline{\dot{v}}(r,t)\cdot\underline{q}(r,t)\right) r\,\mathrm{d}r\mathrm{d}t \tag{7.6.17}$$

for

$$\underline{q} = \begin{pmatrix}\overline{\eta}_i^p(r)\eta_l^{n-1}(t)\\ 0\end{pmatrix}, \quad \begin{pmatrix}0\\ \overline{\eta}_i^p(r)\eta_l^{n-1}(t)\end{pmatrix}, \quad l=0,1,\ldots,n-1, \quad i=0,1,\ldots,p, \tag{7.6.18}$$

to generate our element residual vector of length $2n(p+1)$. For the two test vectors in (7.6.18) these are respectively the $2(in+l)$ and $2(in+l)+1$ entries. The ne residual vectors are assembled to form the global residual vector of length $2n(p(ne)+1)$. The Newton iteration that we are using needs a Jacobian matrix and we get the contributions

to the full matrix element-by-element with each element Jacobian matrix having size $2n(p+1)$-by-$2n(p+1)$. With indexing starting from 0, row $2(in+l)$ corresponds to using the first test vector shown and it obtained by partially differentiating with respect to $c_0, c_1, \ldots, c_{2n(p+1)-1}$. Similarly row $2(in+l)+1$ corresponds to the other test vector shown and with partially differentiating this with respect to $c_0, c_1, \ldots, c_{2n(p+1)-1}$. The kth entry of one of the rows is described by

$$\int_{t_{j-1}}^{t_j}\int_{r_s}^{r_{s+1}} \left(\tilde{a}'(\underline{u}(r,t); \frac{\partial \underline{u}}{\partial c_k},\ \underline{q}(r,t)) + \rho h_0 \frac{\partial \dot{\underline{v}}(r,t)}{\partial c_k} \cdot \underline{q}(r,t) \right) r\,\mathrm{d}r\mathrm{d}t, \tag{7.6.19}$$

where here $\tilde{a}'(\cdot;\cdot,\cdot)$ is the Gâteaux derivative of $\tilde{a}(\cdot;\cdot)$. When $k = 2(\bar{i}n + \bar{l})$ we have

$$\frac{\partial \underline{u}}{\partial c_k} = \begin{pmatrix} \overline{\eta}_{\bar{i}}^p(r)\eta_{\bar{l}+1}^n(t) \\ 0 \end{pmatrix}, \quad \frac{\partial \dot{\underline{v}}}{\partial c_k} = \begin{pmatrix} \overline{\eta}_{\bar{i}}^p(r)\ddot{\eta}_{\bar{l}+1}^n(t) + \overline{\eta}_{\bar{i}}^p(r)\dot{\eta}_{\bar{l}+1}^n(t_{j-1})\dot{P}_n(t) \\ 0 \end{pmatrix},$$

and when $k = 2(\bar{i}n + \bar{l}) + 1$ we have

$$\frac{\partial \underline{u}}{\partial c_k} = \begin{pmatrix} 0 \\ \overline{\eta}_{\bar{i}}^p(r)\eta_{\bar{l}+1}^n(t) \end{pmatrix}, \quad \frac{\partial \dot{\underline{v}}}{\partial c_k} = \begin{pmatrix} 0 \\ \overline{\eta}_{\bar{i}}^p(r)\ddot{\eta}_{\bar{l}+1}^n(t) + \overline{\eta}_{\bar{i}}^p(r)\dot{\eta}_{\bar{l}+1}^n(t_{j-1})\dot{P}_n(t) \end{pmatrix}.$$

When the element Jacobian matrices are assembled we get a global Jacobian matrix with $2n(p(ne)+1)$ rows and with a bandwidth of $2(p+1)n-1$. As well as this degree n in time case being more complicated than the $n=1$ case we note here that when $n>1$ the Jacobian matrix of the nonlinear equations to solve is not symmetric.

7.6.6 Comments about the numerical quadrature

As mentioned a few times, the integrals which appear in (7.6.17) and (7.6.19) are approximated using a quadrature rule and we now give brief details of what this involves. For a mapping from a standard element to our actual element we have

$$\begin{aligned} r(x_1) &= \left(\frac{r_s + r_{s+1}}{2}\right) + \left(\frac{r_{s+1} - r_s}{2}\right) x_1, \quad -1 < x_1 < 1, \\ t(x_2) &= \left(\frac{t_{j-1} + t_j}{2}\right) + \left(\frac{t_j - t_{j-1}}{2}\right) x_2, \quad -1 < x_2 < 1. \end{aligned}$$

Hence for any given integrand $g(r,t)$ we have

$$\int_{t_{j-1}}^{t_j}\int_{r_s}^{r_{s+1}} g(r,t)r\,\mathrm{d}r\mathrm{d}t = \int_{t_{j-1}}^{t_j}\int_{r_s}^{r_{s+1}} g(r(x_1),t(x_2))\left|\frac{dr(x_1)}{dx_1}\right|\left|\frac{dt(x_2)}{dx_2}\right| r(x_1)\,\mathrm{d}x_1\mathrm{d}x_2$$
$$= \left(\frac{r_{s+1}-r_s}{2}\right)\left(\frac{t_j-t_{j-1}}{2}\right)\int_{-1}^{1}\int_{-1}^{1} g(r(x_1),t(x_2))r(x_1)\,\mathrm{d}x_1\mathrm{d}x_2.$$

In our case each integrand g involves a product of terms with each involving $\underline{u}$ (and various derivatives) and also involving the test vector $\underline{q}$. Based on the degree of polynomials that this involves we use Gauss Legendre quadrature of degree n in the t direction and of degree $p+1$ in the r direction. If for a m point Gauss Legendre rule on $(-1,1)$ the points are denoted by ξ_i^m, $i=0,\ldots,m-1$ and the corresponding weights are denoted by w_i^m then the approximation to the integral is given by

$$\left(\frac{r_{s+1}-r_s}{2}\right)\left(\frac{t_j-t_{j-1}}{2}\right)\sum_{i=1}^{n}\sum_{l=1}^{p+1} w_i^n w_l^{p+1} g(r(\xi_l^{p+1}),t(\xi_i^n))r(\xi_l^{p+1}).$$

7.6.7 *How to get $\underline{u}^j(r)$ and $\underline{v}^j(r)$*

Once we have solved the nonlinear equations to determine all the parameters describing $\underline{u}(r,t)$ on $0\le r\le 1$, $t_{j-1}\le t\le t_j$ we need to get the parameters describing $\underline{u}^j(r)$ and $\underline{v}^j(r)$ in a suitable format to be able to consider the next time interval $t_j\le t\le t_{j+1}$. We have already mentioned how to get $\underline{u}^j(r)$ from $\underline{u}(r,t)$ by extracting the appropriate $2(p(ne)+1)$ parameters and specifically these are

$$c_{2(in+n-1)}, c_{2(in+n-1)+1}, \quad i=0,1,\ldots,p(ne). \tag{7.6.20}$$

In the case of $\underline{v}^j(r)$ there is a bit more to do as we have to evaluate at time t_j. As $P_n(t_j)=1$ and using (7.6.13) we have

$$\underline{v}^j(r) = \dot{\underline{u}}(r,t_j) + \underline{\gamma}(r) = \dot{\underline{u}}(r,t_j) + (-1)^n\left(\underline{v}^{j-1}(r) - \dot{\underline{u}}(r,t_{j-1}).\right) \tag{7.6.21}$$

On an element we can write this as

$$\begin{aligned}\underline{v}^j(r) &= (-1)^n\underline{v}^{j-1}(r) + \underline{u}^{j-1}(r)\left(\dot{\eta}_0(t_j) + (-1)^{n+1}\dot{\eta}_0(t_{j-1})\right)\\ &+ \sum_{i=0}^{p}\overline{\eta}_i^p(r)\sum_{l=0}^{n-1}\begin{pmatrix} c_{2(in+l)}\\ c_{2(in+l)+1}\end{pmatrix}\left(\dot{\eta}_{l+1}^n(t_j) + (-1)^{n+1}\dot{\eta}_{l+1}^n(t_{j-1})\right).\end{aligned} \tag{7.6.22}$$

This shows how we get the coefficients in the representation of $\underline{v}^j(r)$ using basis functions given in (7.6.7).

7.6.8 A summary of the steps when solving in $[t_{j-1}, t_j]$

We now summarize the solution process in the time interval $[t_{j-1}, t_j]$ in the finite element scheme with the full set of nonlinear equations described by

$$\underline{f}(\underline{c}) = \underline{0}$$

and with the solution process involving the Newton iteration starting with $\underline{c}^{(0)}$ and then continuing with

$$\underline{c}^{(k+1)} = \underline{c}^{(k)} - J_f(\underline{c}^{(k)})^{-1}\underline{f}(\underline{c}^{(k)}), \quad k = 0, 1, 2, \ldots$$

where $J_f(\underline{c}^{(k)})$ represents the Jacobian matrix corresponding to $\underline{f}(\underline{c}^{(k)})$.

A pseudo code description of the computations is as follows.

For $k = 0, 1, \ldots$ until the maximum number of iterations allowed

For elements $s = 0, 1, \ldots, ne - 1$

We compute quadrature approximations to the integrals in (7.6.17),

$l = 0, \ldots, n-1$, $i = 0, \ldots, p+1$ to get the element residual.

We compute quadrature approximations to the integrals in (7.6.19)

to get the element Jacobian matrix.

We apply boundary conditions when $s = 0$ or $s = ne - 1$.

We assemble the element terms to partly get $\underline{f}(\underline{c}^{(k)})$ and $J_f(\underline{c}^{(k)})$.

End of element loop

We solve a banded system to enable us to get $\underline{c}^{(k+1)}$.

We leave the loop if $\left\|\underline{f}(\underline{c}^{(k)})\right\|$ is sufficiently small.

End of Newton iteration loop.

We extract the parameters which describe $\underline{u}^j(r)$ as in (7.6.20).

We do the computations implied by (7.6.21) to get the parameters which describe $\underline{v}^j(r)$.

7.7 A higher order scheme in time for the dual problem

We keep this section brief and just concentrate on the main differences between what has already been described in section 7.5 when we obtain the dual solution in a time interval when polynomials of any degree in time are used in the approximation.

As before, in the dual problem the unknowns are denoted by $\underline{\psi}(r,t)$ and $\underline{\theta}(r,t)$ and we are solving backward in time. A higher order in time scheme can be described in a similar way to that given earlier to get $\underline{u}(r,t)$ and $\underline{v}(r,t)$ with a few adjustments. To avoid cumbersome notation we use again p and n to denote the degrees of polynomials on each element in space and time respectively for the components of $\underline{\psi}$ and $\underline{\theta}$ and these will need to be at least as high as that used for the finite element displacement $\underline{u}$ and velocity $\underline{v}$. We also again use the notation

$$0 = t_0 < t_1 < \cdots < t_N = T$$

for the time levels and for the values at times t_j and t_{j-1} we let, similar to (7.6.2) and (7.6.3),

$$\underline{\psi}^{j-1}(r) := \underline{\psi}(r,t_{j-1}), \qquad \underline{\theta}^{j-1}(r) := \underline{\theta}(r,t_{j-1}), \tag{7.7.1}$$

$$\underline{\psi}^{j}(r) := \underline{\psi}(r,t_{j}), \qquad \underline{\theta}^{j}(r) := \underline{\theta}(r,t_{j}). \tag{7.7.2}$$

In all the implementations considered so far with the higher order scheme we have used the same space mesh and the same time levels for the dual problem with the higher order schemes as we have used for the problem to get the approximations to $\underline{u}$ and $\underline{v}$. Similar to what was given in the description to get $\underline{u}(r,t)$ and $\underline{v}(r,t)$ we now have the following on an element $r_s \le r \le r_{s+1}$, $t_{j-1} \le t \le t_j$.

$$\underline{\psi}^{j-1}(r), \quad \underline{\theta}^{j-1}(r) \in \text{span}\left\{ \begin{pmatrix} \overline{\eta}_0^p(r) \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ \overline{\eta}_0^p(r) \end{pmatrix}, \ldots, \begin{pmatrix} \overline{\eta}_p^p(r) \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ \overline{\eta}_p^p(r) \end{pmatrix} \right\}, \tag{7.7.3}$$

and

$$\underline{\psi}(r,t) = \underline{\psi}^j(r)\eta_n^n(t) + \sum_{i=0}^{p} \overline{\eta}_i^p(r) \sum_{l=0}^{n-1} \begin{pmatrix} d_{2(in+l)} \\ d_{2(in+l)+1} \end{pmatrix} \eta_l^n(t). \tag{7.7.4}$$

It is written like this as when we are solving backward in time and we are considering this time interval, we already know the solution at time t_j and we are attempting to determine the other parameters associated with the time interval.

As it was described in section 6.6 the equations to determine $\underline{\psi}(r,t)$ and $\underline{\theta}(r,t)$ in

$0 \le r \le 1$, $t_{j-1} \le t \le t_j$ are

$$-\rho h_0 \int_{t_{j-1}}^{t_j}\int_0^1 \left(\underline{\theta}(r,t)+\dot{\underline{\psi}}(r,t)\right)\cdot \underline{q}(r,t) r\,\mathrm{d}r\mathrm{d}t = \int_{t_{j-1}}^{t_j}\int_0^1 \underline{q}(r,t)\cdot \underline{J}_{\underline{\beta}} r\,\mathrm{d}r\mathrm{d}t, \tag{7.7.5}$$

$$\int_{t_{j-1}}^{t_j}\int_0^1 \left(\tilde{a}'(\underline{u}(r,t);\underline{q}(r,t),\underline{\psi}(r,t)) - \rho h_0 \dot{\underline{\theta}}(r,t)\cdot \underline{q}(r,t)\right) r\,\mathrm{d}r\mathrm{d}t \tag{7.7.6}$$

$$= \int_{t_{j-1}}^{t_j}\int_0^1 \left(\underline{q}(r,t)\cdot \underline{J}_{\underline{\alpha}} + \underline{q}'(r,t)\cdot \underline{J}_{\underline{\alpha}'}\right) r\,\mathrm{d}r\mathrm{d}t,$$

where $\tilde{a}'(\cdot;\cdot,\cdot)$ represents the Gâteaux derivative of $\tilde{a}'(\cdot;\cdot)$.
In each case it is for all $\underline{q}$ in the following span of the functions

$$\begin{pmatrix} \overline{\eta}_i^p(r)\eta_l^{n-1}(t) \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ \overline{\eta}_i^p(r)\eta_l^{n-1}(t) \end{pmatrix}, \qquad l = 0,1,\ldots,n-1, \quad i = 0,1,\ldots,p. \tag{7.7.7}$$

For the first equation (7.7.5), we restrict to quantities of interest for which $\underline{\theta}+\dot{\underline{\psi}} = \underline{0}$ and this now gives for our approximations that $\underline{\theta}$ and $\dot{\underline{\psi}}$ are related by

$$\underline{\theta}(r,t) + \dot{\underline{\psi}}(r,t) = \underline{\gamma}(r) P_n(t) \tag{7.7.8}$$

where $P_n(t)$ is the Legendre polynomial of degree n on $[t_{j-1}, t_j]$. As $P_n(t_j) = 1$ it follows that

$$\underline{\gamma}(r) = \underline{\theta}^j(r) + \dot{\underline{\psi}}(r,t_j). \tag{7.7.9}$$

For the second equation (7.7.6), we need $\dot{\underline{\theta}}(r,t)$ and this is given by

$$\dot{\underline{\theta}}(r,t) = -\ddot{\underline{\psi}}(r,t) + \underline{\gamma}(r)\dot{P}_n(t). \tag{7.7.10}$$

On an element this can be represented as follows.

$$\begin{aligned}
\dot{\underline{\theta}}(r,t) &= -\underline{\psi}^j(r)\ddot{\eta}_n^n(t) - \sum_{i=0}^{p}\overline{\eta}_i^p(r)\sum_{l=0}^{n-1}\begin{pmatrix} d_{2(in+l)} \\ d_{2(in+l)+1}\end{pmatrix}\ddot{\eta}_l^n(t) \\
&\quad + \left(\underline{\theta}^j(r) + \underline{\psi}^j(r)\dot{\eta}_n^n(t_j) + \sum_{i=0}^{p}\overline{\eta}_i^p(r)\sum_{l=0}^{n-1}\begin{pmatrix} d_{2(in+l)} \\ d_{2(in+l)+1}\end{pmatrix}\dot{\eta}_l^n(t_j)\right)\dot{P}_n(t) \\
&= -\sum_{i=0}^{p}\overline{\eta}_i^p(r)\sum_{l=1}^{n-1}\begin{pmatrix} d_{2(in+l)} \\ d_{2(in+l)+1}\end{pmatrix}\ddot{\eta}_l^n(t) \\
&\quad + \left(\underline{\theta}^j(r) + \underline{\psi}^j(r)\dot{\eta}_n^n(t_j) + \sum_{i=0}^{p}\overline{\eta}_i^p(r)\sum_{l=0}^{n-1}\begin{pmatrix} d_{2(in+l)} \\ d_{2(in+l)+1}\end{pmatrix}\dot{\eta}_l^n(t_j)\right)\dot{P}_n(t)
\end{aligned}$$

where the last version is because the second time derivative of the first and last basis

functions is zero.

Once a solution is obtained in the time interval $t_{j-1} \le t \le t_j$ we get the coefficients

$$d_{2in}\,, d_{2in+1}, \quad i = 0, 1, \ldots, p(ne).$$

needed for the function $\underline{\psi}^{j-1}(r)$ with respect to the basis

$$\text{span}\left\{ \begin{pmatrix} \overline{\eta}_0^p(r) \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ \overline{\eta}_0^p(r) \end{pmatrix}, \ldots, \begin{pmatrix} \overline{\eta}_p^p(r) \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ \overline{\eta}_p^p(r) \end{pmatrix} \right\}.$$

For the coefficients of the function $\underline{\theta}^{j-1}(r) = \underline{\theta}(r, t_{j-1})$ we have

$$\begin{aligned}
\underline{\theta}^{j-1}(r) &= (-1)^n \underline{\gamma}(r) - \dot{\underline{\psi}}(r, t_{j-1}) \\
&= (-1)^n (\underline{\theta}^j(r) + \dot{\underline{\psi}}(r, t_j)) - \dot{\underline{\psi}}(r, t_{j-1}) \\
&= (-1)^n \underline{\theta}^j(r) + ((-1)^n \dot{\underline{\psi}}(r, t_j) - \dot{\underline{\psi}}(r, t_{j-1})) \\
&= (-1)^n \underline{\theta}^j(r) + \underline{\psi}^j(r)((-1)^n \dot{\eta}_n^n(t_j) - \dot{\eta}_n^n(t_{j-1})) \qquad (7.7.11) \\
&\quad + \sum_{i=0}^{p} \overline{\eta}_i^p(r) \sum_{l=0}^{n-1} \begin{pmatrix} d_{2(in+l)} \\ d_{2(in+l)+1} \end{pmatrix} ((-1)^n \dot{\eta}_l^n(t_j) - \dot{\eta}_l^n(t_{j-1})). \qquad (7.7.12)
\end{aligned}$$

From this representation we get the coefficients of the function $\underline{\theta}^{j-1}(r)$.

8. RESULTS WITH A HYPERELASTIC AXI-SYMMETRIC MEMBRANE MODEL

8.1 Introduction

In this chapter we present results to test the theory described in chapter 6 to show how the error estimate values for a quantity of interest (QoI) obtained via the solution of a dual problem compare with the actual error, when this is possible. Here, the term actual error will be in practice involve using the most accurate estimate of the QoI to serve as the exact value. We also describe how we can use the error estimate expression to help us to adaptively refine, in some cases, so that we have a goal orientated technique to get an estimate of a QoI to a desired accuracy. This is straightforward to do in the quasi-static case when the computational effort is never too great but in the dynamic case it is only used to help guide us towards a more accurate result in terms of whether we should refine in both space and time or in just one of these at each refinement step.

The order in which the results are presented corresponds to the order the theory was given in chapter 6. We consider first the quasi-static case and then the dynamic case. The section on the dynamic case is split into the parts when just the basic scheme in time is used and when the higher order in time is used.

8.2 The quasi-static problem

We start with the quasi-static inflation problem, where the time-dependence is only through the time-dependent pressure $P(t)$. First we give the desired functionals J that we want to approximate, we give the expression for the Gâteaux derivative J' needed in a dual problem and then give some numerical results with different numbers of elements and with different degrees for the piecewise polynomial approximations to the membrane displacement $\underline{u}(r)$.

8.2.1 The desired quantities of interest for the quasi-static case

First we will describe the desired quantities of interest $J(\underline{u})$ which we use in our results in this section and we also give details of the Gâteaux derivative $J'(\underline{u};\underline{\alpha})$ which is needed in the dual problem. To repeat what was given in (7.2.2) we consider functionals of the form

$$J(\underline{u}) = \int_{\Omega^*} \{\text{expressions in } \underline{u} \text{ and } \underline{u}'\}\ r dr,$$

where Ω^* is Ω or it is part of Ω, which gives a Gâteaux derivative of the form

$$J'(\underline{u};\alpha) = \int_0^1 (\underline{\alpha}\cdot\underline{J}_\alpha + \underline{\alpha}'\cdot\underline{J}'_\alpha) r\, dr, \tag{8.2.1}$$

$$= \int_0^1 (\alpha_1 J_{\alpha_1} + \alpha_3 J_{\alpha_3} + \alpha'_1 J_{\alpha'_1} + \alpha'_3 J_{\alpha_3}) r\, dr. \tag{8.2.2}$$

We denote the particular functionals that we consider as J_1 and J_2 in the following.

The thickness stretch ratio over a part or all of the domain

Let

$$J_1(\underline{u}) = \int_0^b \lambda(r) r dr = \int_0^b \frac{1}{\lambda_1(r)\lambda_2(r)} r dr. \tag{8.2.3}$$

with λ being the stretch ratio through the thickness. If we divide this quantity by $b^2/2$ then we get the average of λ over the part of the domain considered. In the results we consider this case when $b = 1$, which is the entire domain, and when $b = 1/8$. When $b = 1/8$ we always use meshes with $r = b$ being one of the nodes.

To get the Gâteaux derivative first consider the Gâteaux derivative of the integrand λ which as a first step involves

$$\lambda'(\underline{u};\underline{\alpha}) = -(\lambda_1\lambda_2)^{-2}(\lambda'_1(\underline{u};\underline{\alpha})\lambda_2 + \lambda_1\lambda_2(\underline{u};\underline{\alpha})). \tag{8.2.4}$$

Now as

$$\lambda_1^2 = (1+u'_1)^2 + {u'_3}^2 \quad \text{and} \quad \lambda_2 = 1 + \frac{u_1}{r}$$

we get

$$\lambda'_1 = \frac{1}{\lambda_1}\left((1+u'_1)\alpha'_1 + u'_3\alpha'_3\right) \quad \text{and} \quad \lambda'_2 = \frac{\alpha_1}{r}. \tag{8.2.5}$$

Thus

$$J_1'(\underline{u};\underline{\alpha}) = \int_0^b -(\lambda_1\lambda_2)^{-2}\left(\lambda_1\frac{\alpha_1}{r} + \frac{\lambda_2}{\lambda_1}\left((1+u_1')\alpha_1' + u_3'\alpha_3'\right)\right) r dr. \tag{8.2.6}$$

This is of the form given in (8.2.2) with

$$J_{\alpha_1} = -(\lambda_1\lambda_2)^{-2}\frac{\lambda_1}{r}, \qquad J_{\alpha_3} = 0, \tag{8.2.7}$$

$$J_{\alpha_1'} = -(\lambda_1\lambda_2)^{-2}\frac{\lambda_2}{\lambda_1}(1+u_1'), \qquad J_{\alpha_3'} = -(\lambda_1\lambda_2)^{-2}\frac{\lambda_2}{\lambda_1}u_3'. \tag{8.2.8}$$

The potential energy of the deformed membrane

Let

$$J_2(\underline{u}) = \Psi(\underline{u}) = \int_0^1 \tilde{\Psi}(r) r\, dr \tag{8.2.9}$$

where

$$\tilde{\Psi}(r) = h_0 W - \frac{P}{3}\left(1 + \frac{u_1(r)}{r}\right)\left(-(r+u_1(r))u_3'(r) + (1+u_1'(r))u_3(r)\right), \tag{8.2.10}$$

with W being the strain energy function given by a hyperelastic model.

Now, the Gâteaux derivatives of this functional is such that $J_2'(\underline{u};\underline{\alpha}) = A_Q(\underline{u};\underline{\alpha})$, as in the non-axisymmetric case, which we show here, and we give some details here as in the dual problem we do need the expressions for J_{α_1}, J_{α_3}, $J_{\alpha_1'}$ and $J_{\alpha_3'}$.

Let

$$g(\underline{u}) = \left(1 + \frac{u_1}{r}\right)\left(-(r+u_1)u_3' + (1+u_1')u_3\right)$$

so that

$$\tilde{\Psi} = h_0 W - \frac{P}{3}g.$$

By the chain rule the Gâteaux derivative of W is

$$\begin{aligned} W'(\underline{u};\underline{\alpha}) &= W_1\lambda_1'(\underline{u};\underline{\alpha}) + W_2\lambda_2'(\underline{u};\underline{\alpha}) \\ &= \frac{W_1}{\lambda_1}((1+u_1')\alpha_1' + u_3'\alpha_3') + W_2\frac{\alpha_1}{r} \end{aligned} \tag{8.2.11}$$

where we have used the expressions for λ_1' and λ_2' given in (8.2.5). For the term involving

g the product rule gives

$$\begin{aligned} g'(\underline{u};\underline{\alpha}) &= \frac{\alpha_1}{r}\left(-(r+u_1)u_3' + (1+u_1')u_3\right) \\ &\quad + \left(1+\frac{u_1}{r}\right)\left(-\alpha_1 u_3' - (r+u_1)\alpha_3' + \alpha_1' u_3 + (1+u_1')\alpha_3\right) \quad (8.2.12) \\ &= \frac{\alpha_1}{r}\left(-2(r+u_1)u_3' + (1+u_1')u_3\right) \\ &\quad + \left(1+\frac{u_1}{r}\right)\left(-(r+u_1)\alpha_3' + \alpha_1' u_3 + (1+u_1')\alpha_3\right). \quad (8.2.13) \end{aligned}$$

If we compare with (6.3.6) and (6.3.9) then we see that $\tilde{\Psi}'(\underline{u};\underline{\alpha}) = h_0 a_1(\underline{u};\underline{\alpha}) - (P/3)a_2(\underline{u};\underline{\alpha})$ to verify that $J'(\underline{u};\underline{\alpha}) = A_Q(\underline{u};\underline{\alpha})$. For the expressions needed here we have

$$\begin{aligned} J_{\alpha_1} &= \frac{h_0 W_2}{r} - \frac{P}{3r}\left(u_3(1+u_1') - 2(r+u_1)u_3'\right), \\ J_{\alpha_3} &= -\frac{P}{3}\left(1+\frac{u_1}{r}\right)(1+u_1'), \\ J_{\alpha_1'} &= h_0 W_1\left(\frac{1+u_1'}{\lambda_1}\right) - \frac{P}{3r}\left(1+\frac{u_1}{r}\right)u_3, \\ J_{\alpha_3'} &= h_0 W_1\left(\frac{u_3'}{\lambda_1}\right) + \frac{P}{3r}\left(1+\frac{u_1}{r}\right)(r+u_1). \end{aligned}$$

Note: As the results only depend on the ratio P/h_0, where P is the time-dependent applied pressure and h_0 is the undeformed thickness, we can take $h_0 = 1$ in the computations and just report the value of P when indicating a particular solution.

8.2.2 The goal-oriented adaptive refinement technique for the quasi-static case

In this section, we present the *goal-oriented* adaptive refinement procedure, which is used in this axisymmetric quasi-static case. This technique is based on the local error contributions that we get when we consider a term of the form $-A(\underline{u}_h;\underline{\tilde{\psi}}_h)$ where $\underline{\tilde{\psi}}_h$ denotes one of the dual solutions that we consider. With V_h denoting the finite element space that we use to get $\underline{u}_h$ and with $\bar{V}_h$ being the larger finite element space that we use when we solve the approximate dual problem, we consider the following two possibilities for $\underline{\tilde{\psi}}_h$ when results are presented. When $\underline{u}_h$ is the data in the dual problem we let $\underline{\psi}_h \in \bar{V}_h$ be such that

$$A'(\underline{u}_h;\underline{\alpha},\underline{\psi}_h) = J'(\underline{u}_h;\underline{\alpha}) \quad \forall \underline{\alpha} \in \bar{V}_h. \quad (8.2.14)$$

When a better approximation $\underline{u}_h^b$ is available, e.g. the finite element solution for the displacement using the space $\bar{V}_h$, we let

$$\underline{u}_h^m = \frac{1}{2}(\underline{u}_h + \underline{u}_h^b) \tag{8.2.15}$$

and we let $\underline{\psi}_h^m \in \bar{V}_h$ be such that

$$A'(\underline{u}_h^m; \underline{\alpha}, \underline{\psi}_h^m) = J'(\underline{u}_h^m; \underline{\alpha}) \quad \forall \underline{\alpha} \in \bar{V}_h. \tag{8.2.16}$$

From what was discussed in chapter 4, we expect that the estimate $-A(\underline{u}_h; \underline{\psi}_h^m)$ to be asymptotically exact in estimating the actual error and we expect that the estimate $-A(\underline{u}_h; \underline{\psi}_h)$ just to be consistent in estimating the actual error. The examples that we give in section 8.2.3 support this.

Before considering the values that we get we describe here how to use a quantity of the form $-A(\underline{u}_h; \tilde{\underline{\psi}}_h)$ to help determine which element to refine, and by how much to refine, with the aim being to get to a desired level of accuracy. From the theory given in chapter 4 we have the following

$$\begin{aligned} J(\underline{u}) - J(\underline{u}_h) &\approx -A(\underline{u}_h; \tilde{\underline{\psi}}_h) \\ &= -A\left(\underline{u}_h; \tilde{\underline{\psi}}_h - \underline{\psi}_{h_I}\right) \quad \forall \underline{\psi}_{h_I} \in V_h \\ &= -\sum_{k=1}^{ne} A\left(\underline{u}_h; \tilde{\underline{\psi}} - \underline{\psi}_{h_I}\right)_k, \quad \forall \underline{\psi}_{h_I} \in V_h \end{aligned}$$

where $A(\cdot;\cdot)_k$ means the expression for $A(\cdot;\cdot)$ but with the integration only over the k^{th} element Ω_k. As we can subtract from $\tilde{\underline{\psi}}_h$ any function in V_h we can do this by subtracting the interpolant to $\tilde{\underline{\psi}}_h$ from V_h and in the case of higher degree polynomials and a basis constructed using Legendre polynomials we can have $\underline{\psi}_{h_I}$ such that $\tilde{\underline{\psi}}_h - \underline{\psi}_{h_I}$ has no lower degree polynomial terms. Now, the quantities

$$A\left(\underline{u}_h; \tilde{\underline{\psi}}_h - \underline{\psi}_{h_I}\right)_k, \quad k = 1, \ldots, ne, \tag{8.2.17}$$

are our local error estimators and they give an indication of the contribution to the error from each element. We make use of this observation to determine how to refine the mesh in an economical way in order to compute $J(\underline{u}_h)$ which will eventually give us

$$|J(\underline{u}) - J(\underline{u}_h)| < \text{tol}$$

where tol is our desired accuracy. In terms of the mesh itself, our aim is to ideally get to

the stage when

$$\left|A\left(\underline{u}_h;\underline{\psi}-\underline{\psi}_{h_I}\right)_k\right| \approx \frac{\text{tol}}{ne}, \quad \text{for } k=1,\ldots,ne. \tag{8.2.18}$$

If the error estimate for the k^{th} element is larger than tol/ne then we need to refine the k^{th} element. Now, for the element refinement we use the following procedure which helps us to reach our goal very fast in terms of the computational costs.

Element refinement procedure

For each k^{th} element to determine by how much it needs to be divided we consider what is needed if we are in the asymptotic convergence range. For example, if we are using degree p polynomials in the approximation of the components of $\underline{u}$ then dividing the element into q_k equal elements should decrease the error estimate corresponding to the region of the k^{th} element by a factor of about q_k^{2p}. This suggest that we should select q_k so that

$$q_k^{2p} \approx \frac{\left|A\left(\underline{u}_h;\underline{\psi}-\underline{\psi}_{h_I}\right)_k\right|}{\text{tol}/ne} > 1. \tag{8.2.19}$$

It is usually beneficial to take a slightly larger value for q_k than this as otherwise the strategy of aiming to almost exactly get to the required accuracy often results in a new mesh which has an error which is slightly larger than the tol. Now if the value obtained for q_k, calculated by (8.2.19) is very large then we replace it by some fixed value with the knock on the effect that more than one mesh refinement will probably be needed before the accuracy of tol is reached. We give this next as an algorithm.

Algorithm

Step 1 Choose a mesh size h and calculate the approximating finite element solution $\underline{u}_h$. If $\underline{\psi}_h^m$ is required then we also need to solve a finite element problem to get $\underline{u}_h^b$ from which we get $\underline{u}_h^m = (\underline{u}_h + \underline{u}_h^b)/2$.

Step 2 Solve a dual problem to get $\underline{\tilde{\psi}}_h$ which is $\underline{\psi}_h$ or $\underline{\psi}_h^m$.

Step 3 Compute the error estimates $A\left(\underline{u}_h;\underline{\tilde{\psi}}_h-\underline{\psi}_{h_I}\right)_k$ for $k=1,\ldots,ne$.

Step 4 Check for accuracy. If

$$\left|A\left(\underline{u}_h;\underline{\tilde{\psi}}_h-\underline{\psi}_{h_I}\right)_k\right| \leq \frac{\text{tol}}{ne}$$

then set $q_k = 1$ and go to the next value of k. Otherwise refine the k^{th} element as

follows. Compute

$$q_k = \text{ceil}\left(\frac{1.05\left|A\left(\underline{u}_h; \tilde{\underline{\psi}}_h - \underline{\psi}_{h_I}\right)_k\right|}{\text{tol/ne}}\right)^{1/(2p)}$$

where p represents the degree of the piecewise polynomials in the finite element approximation $\underline{u}_h$. If $q_k > 16$ then re-set $q_k = 16$.

Step 5 Construct a new mesh by subdividing into $q_k + 1$ elements each k^{th} element that needs refinement, based on error estimators.

Step 6 Replace the old mesh with the new mesh.

Step 7 Repeat the procedure until we obtain

$$\left|\sum_{k=1}^{ne} A\left(\underline{u}_h; \tilde{\underline{\psi}}_h - \underline{\psi}_{h_I}\right)_k\right| \leq \text{tol}.$$

Note: The factor 1.05 and the upper error bound of 16 that we used above were chosen based on the computational results we got during the implementation of the mesh refinement. Other values can also be considered.

8.2.3 Numerical examples for the quasi-static case

We present here a fairly comprehensive set of results to demonstrate numerically several aspects of the theory in the case of the functionals J_1 and J_2 described in section 8.2.1. We consider piecewise polynomial approximations of degree $p = 1$, $p = 2$ and $p = 3$, we give the estimates of the error $J(\underline{u}) - J(\underline{u}_h)$ using $-a(\underline{u}_h; \underline{\psi}_h)$ and with using $-a(\underline{u}_h; \underline{\psi}_h^m)$ where $\underline{\psi}_h$ and $\underline{\psi}_h^m$ are indicated in (8.2.14)–(8.2.16), and we give results using the adaptive refinement algorithm given at the end of section 8.2.2. In each of the results given, when the finite element space V_h involves piecewise polynomials of degree p to get $\underline{u}_h$ the space $\bar{V}_h$ used to get $\underline{\psi}_h$ and $\underline{\psi}_h^m$ involves piecewise polynomials of degree $p + 1$.

There are two deformations that we consider and these correspond to the outer profile in figures 8.1(a) and 8.1(b). As the deformations just depend on the ratio P/h_0 we take $h_0 = 1$. In figure 8.1(a) where the deformation is not too large the Jones-Treloar model is used and the outer profile is at pressure $P = 0.3$. In figure 8.1(b) where the deformation

is much larger, the Mooney-Rivlin model is used with the strain energy function given by

$$W = \frac{1}{2}\left(\lambda_1^2 + \lambda_2^2 + \lambda_3^2 - 3\right) + \frac{0.1}{2}\left(\lambda_1^{-2} + \lambda_2^{-2} + \lambda_3^{-2} - 3\right) \tag{8.2.20}$$

where, as before, λ_1 and λ_2 denote the principal stretch ratios with $\lambda_3 = 1/(\lambda_1\lambda_2)$. The other profiles shown in the figures show some of the intermediate deformations obtained with lower pressures. The other profile in figure 8.1(a) is at pressure $P = 0.15$ whilst in figure 8.1(b) the other profiles are for pressures at the equally spaced values $P = 0.3, 0.6, \ldots, 2.7$. The quantities of interest that we wish to estimate are as follows. In the case of figure 8.1(a) it is $J_1(\underline{u})$ when $b = 1$, i.e. an integral of the thickness over the entire domain, it is $J_1(\underline{u})$ when $b = 1/8$, i.e. an integral of the thickness over a region near the pole, and it is the potential energy given by $J_2(\underline{u})$. In the case of figure 8.1(b) it is also these 3 cases. To get a value which we can treat as the exact value for the comparisons just requires taking sufficiently many elements and/or a sufficiently high piecewise polynomial approximation and it is more than sufficient to take here ne= 64 elements and a high degree of $p = 8$ which gives the following values.

Tab. 8.2.1: Exact values for the Jones Treloar model, $P = 0.3$ as in figure 8.1(a)

Description of $J(\underline{u})$	The value of $J(\underline{u})$
$J_1(\underline{u})$ with $b = 1$	3.25304427515041e-01
$J_1(\underline{u})$ with $b = 1/8$	4.92523975104043e-03
$J_2(\underline{u})$	2.75959154012544e-02

Tab. 8.2.2: Exact values for the Mooney Rivlin model, $P = 3$ as in figure 8.1(b)

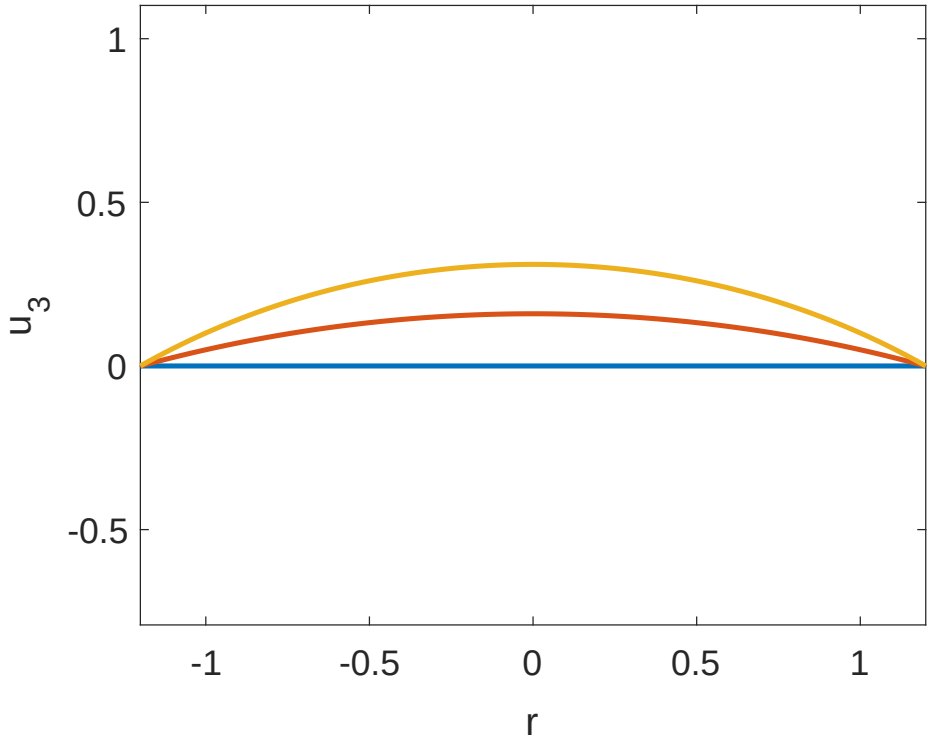
Description of $J(\underline{u})$	The value of $J(\underline{u})$
$J_1(\underline{u})$ with $b = 1$	-1.96972631920304e+00
$J_1(\underline{u})$ with $b = 1/8$	4.01069441687288e-02
$J_2(\underline{u})$	3.31383507600083e-04

As we explained in chapter 4, we expect the estimate $-a(\underline{u}_h; \underline{\psi}_h)$ to be consistent with $J(\underline{u}) - J(\underline{u}_h)$ and we expect the estimate $-a(\underline{u}_h; \underline{\psi}_h^m)$ to be asymptotically exact, i.e. we expect that

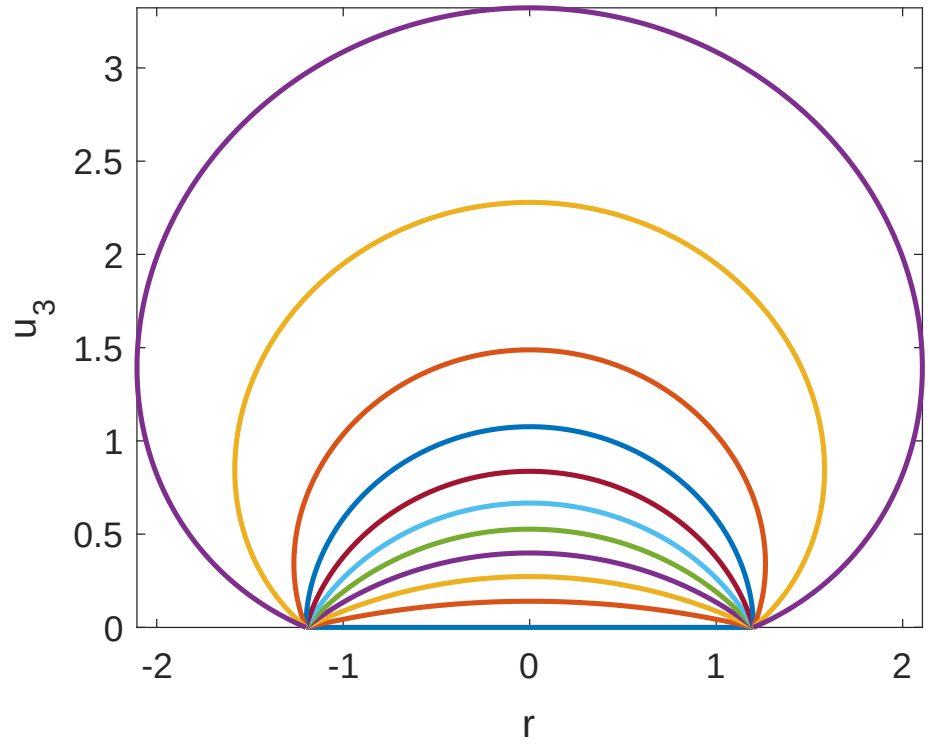
$$\frac{-a(\underline{u}_h; \underline{\psi}_h^m)}{J(\underline{u}) - J(\underline{u}_h)} \to 1 \quad \text{as } h \to 0. \tag{8.2.21}$$

To test for consistency and asymptotic exactness in all the tables we have a column showing the value

$$\frac{-a(\underline{u}_h; \tilde{\underline{\psi}}_h)}{J(\underline{u}) - J(\underline{u}_h)} - 1, \tag{8.2.22}$$



(a) Jones Treloar model with $P = 0.3$



(b) Mooney Rivlin model with $P = 3$

Fig. 8.1: The deformed profiles with modest deformation in the top figure and a large deformation in the bottom figure.

where $\underline{\tilde{\psi}}_h$ is either of $\underline{\psi}_h$ or $\underline{\psi}_h^m$. It is values tending to 0 which are the asymptotically exact case. We refer to this as the asymptotic exactness value in the tables.

The order of the tables that we give corresponds to which part of the theory we wish to illustrate most. Firstly we demonstrate that for each QoI functional J considered

$$J(\underline{u}) - J(\underline{u}_h) = \mathcal{O}(h^{2p}), \tag{8.2.23}$$

the estimate of the error given by $-a(\underline{u}_h; \underline{\psi}_h)$ is consistent with the actual error and the estimate of the error given by $-a(\underline{u}_h; \underline{\psi}_h^m)$ is asymptotically exact. We then show results when we adaptively refine to reach a specified accuracy using one of two refinement steps and we also compare the exact error and the estimates of this in these adaptive refinement cases.

Uniform refinement when the degree $p = 1$

The next two tables are for the functional J_1 when $b = 1$ and the deformation shown in figure 8.1(a). In both cases the exact error column means values $J_1(\underline{u}) - J_1(\underline{u}_h)$ and the column ratio has the ratio of successive error estimates. The asymptotic exactness value is as given in (8.2.22).

Tab. 8.2.3: The error estimate is $-a(\underline{u}_h; \underline{\psi}_h^m)$.

ne	Exact error	Estimates with linears	Asymptotic exactness value	Ratios
10	-4.92587e-005	-4.92492e-005	-1.932016e-004	
20	-1.23625e-005	-1.23616e-005	-7.505904e-005	3.984047
40	-3.09406e-006	-3.09400e-006	-1.876031e-005	3.995346
80	-7.73762e-007	-7.73758e-007	-4.689929e-006	3.998666

Tab. 8.2.4: The error estimate is $-a(\underline{u}_h; \underline{\psi}_h)$.

ne	Exact error	Estimates with linears	Asymptotic exactness value	Ratios
10	-4.92587e-005	-1.64796e-005	-6.654490e-001	
20	-1.23625e-005	-4.12561e-006	-6.662810e-001	3.994464
40	-3.09406e-006	-1.03192e-006	-6.664837e-001	3.997994
80	-7.73762e-007	-2.58022e-007	-6.665360e-001	3.999349

The next two tables are for the functional J_1 when $b = 1/8$ and the deformation shown in figure 8.1(a). In both cases the exact error column means values $J_1(\underline{u}) - J_1(\underline{u}_h)$ and the column ratio is the ratio of successive error estimates. The asymptotic exactness value is as given in (8.2.22).

Tab. 8.2.5: The error estimate is $-a(\underline{u}_h; \underline{\psi}_h^m)$.

ne	Exact error	Estimates with linears	Asymptotic exactness value	Ratios
8	6.85176e-006	6.89246e-006	5.939109e-003	
16	1.61568e-006	1.62114e-006	3.375673e-003	4.251613
32	3.96799e-007	3.97148e-007	8.790208e-004	4.081954
64	9.86544e-008	9.86764e-008	2.223581e-004	4.024752

Tab. 8.2.6: The error estimate is $-a(\underline{u}_h; \underline{\psi}_h)$.

ne	Exact error	Estimates with linears	Asymptotic exactness value	Ratios
8	6.85176e-006	7.65920e-006	1.178444e-001	
16	1.61568e-006	1.84257e-006	1.404269e-001	4.156803
32	3.96799e-007	4.54915e-007	1.464622e-001	4.050361
64	9.86544e-008	1.13304e-007	1.484895e-001	4.014995

The next two tables are for the functional J_2 and the deformation shown in figure 8.1(a). In both cases the exact error column means values $J_2(\underline{u}) - J_2(\underline{u}_h)$ and the column ratio has the ratio of successive error estimates. The asymptotic exactness value is as given in (8.2.22).

Tab. 8.2.7: The error estimate is $-a(\underline{u}_h; \underline{\psi}_h^m)$.

ne	Exact error	Estimates with linears	Asymptotic exactness value	Ratios
10	-3.25065e-005	-3.25064e-005	-3.451943e-006	
20	-8.16542e-006	-8.16541e-006	-7.388602e-007	3.980988
40	-2.04420e-006	-2.04420e-006	-1.905278e-007	3.994428
80	-5.11255e-007	-5.11255e-007	-4.857968e-008	3.998396

Tab. 8.2.8: The error estimate is $-a(\underline{u}_h; \underline{\psi}_h)$.

ne	Exact error	Estimates with linears	Asymptotic exactness value	Ratios
10	-3.25065e-005	-6.50109e-005	9.999353e-001	
20	-8.16542e-006	-1.63307e-005	9.999829e-001	3.980901
40	-2.04420e-006	-4.08840e-006	9.999956e-001	3.994399
80	-5.11255e-007	-1.02251e-006	9.999989e-001	3.998396

All the tables indicate that when we double the number of elements, which divides h by 2, we decrease the error by about $2^2 = 4$. The tables also show that we get asymptotic exactness when the estimate is $-a(\underline{u}_h; \underline{\psi}_h^m)$ but that the estimate $-a(\underline{u}_h; \underline{\psi}_h)$ is only consistent with the actual error.

Uniform refinement when the degree $p = 2$

The next two tables are for the functional J_1 when $b = 1$ and the deformation shown in figure 8.1(a). In both cases the exact error column means values $J_1(\underline{u}) - J_1(\underline{u}_h)$ and the column ratio has the ratio of successive error estimates. The asymptotic exactness value is as given in (8.2.22).

Tab. 8.2.9: The error estimate is $-a(\underline{u}_h; \underline{\psi}_h^m)$.

ne	Exact error	Estimates with quadratics	Asymptotic exactness value	Ratios
10	-2.03137e-008	-2.03135e-008	-9.465689e-006	
20	-1.27263e-009	-1.27263e-009	-3.156862e-006	15.961827
40	-7.95932e-011	-7.95934e-011	2.991101e-006	15.989140
80	-4.97535e-012	-4.97554e-012	3.663964e-005	15.996937

Tab. 8.2.10: The error estimate is $-a(\underline{u}_h; \underline{\psi}_h)$.

ne	Exact error	Estimates with quadratics	Asymptotic exactness value	Ratios
10	-2.03137e-008	-2.27888e-008	1.218432e-001	
20	-1.27263e-009	-1.42773e-009	1.218696e-001	15.961561
40	-7.95932e-011	-8.92943e-011	1.218844e-001	15.989038
80	-4.97535e-012	-5.58197e-012	1.219250e-001	15.996915

The next two tables are for the functional J_1 when $b = 1/8$ and the deformation shown in figure 8.1(a). In both cases the exact error column means values $J_1(\underline{u}) - J_1(\underline{u}_h)$ and the column ratio is the ratio of successive error estimates. The asymptotic exactness value is as given in (8.2.22).

Tab. 8.2.11: The error estimate is $-a(\underline{u}_h; \underline{\psi}_h^m)$.

ne	Exact error	Estimates with quadratics	Asymptotic exactness value	Ratios
8	-8.06079e-008	-8.05900e-008	-2.216413e-004	
16	-5.41971e-009	-5.41974e-009	5.068628e-006	14.869717
32	-3.46655e-010	-3.46656e-010	1.249158e-006	15.634346
64	-2.18027e-011	-2.18027e-011	6.977139e-007	15.899682

Tab. 8.2.12: The error estimate is $-a(\underline{u}_h; \underline{\psi}_h)$.

ne	Exact error	Estimates with quadratics	Asymptotic exactness value	Ratios
8	-8.06079e-008	-8.08974e-008	3.591009e-003	
16	-5.41971e-009	-5.43844e-009	3.457056e-003	14.875111
32	-3.46655e-010	-3.47817e-010	3.351171e-003	15.635923
64	-2.18027e-011	-2.18752e-011	3.324324e-003	15.90006

The next two tables are for the functional J_2 and the deformation shown in figure 8.1(a). In both cases the exact error column means values $J_2(\underline{u}) - J_2(\underline{u}_h)$ and the column ratio has the ratio of successive error estimates. The asymptotic exactness value is as given in (8.2.22).

Tab. 8.2.13: The error estimate is $-a(\underline{u}_h; \underline{\psi}_h^m)$.

ne	Exact error	Estimates with quadratics	Asymptotic exactness value	Ratios
10	-4.01198e-009	-4.01201e-009	9.311140e-006	
20	-2.51143e-010	-2.51144e-010	2.308432e-006	15.974938
40	-1.57042e-011	-1.57042e-011	-1.000163e-006	15.992154
80	-9.81590e-013	-9.81657e-013	6.828389e-005	15.9976448

Tab. 8.2.14: The error estimate is $-a(\underline{u}_h; \underline{\psi}_h)$.

ne	Exact error	Estimates with quadratics	Asymptotic exactness value	Ratios
10	-4.01198e-009	-8.02435e-009	1.000100e+000	
20	-2.51143e-010	-5.02293e-010	1.000025e+000	15.975437
40	-1.57042e-011	-3.14085e-011	1.000003e+000	15.992263
80	-9.81590e-013	-1.96331e-012	1.000138e+000	15.997728

All the tables indicate that when we double the number of elements, which divides h by 2, we decrease the error by about $2^4 = 16$. The tables also show that we get asymptotic exactness when the estimate is $-a(\underline{u}_h; \underline{\psi}_h^m)$ but that the estimate $-a(\underline{u}_h; \underline{\psi}_h)$ is only consistent with the actual error. The results in table 8.2.13 indicate that we are getting close to what is possible to show with usual floating point arithmetic as J_2 is of order 10^{-4}, the exact error is about 10^{-12} when ne= 80 (i.e. about 8 digits of accuracy) and the asymptotic exactness value is about 10^{-5}. This is the most likely explanation as to why the asymptotic exactness value with ne= 80 is larger than with ne= 40 although all the values in that column are small.

The next three tables are for the deformation shown in figure 8.1(b) and in all cases the error estimate is $-a(\underline{u}_h; \underline{\psi}_h^m)$. In all cases the exact error column means values $J_1(\underline{u}) - J_1(\underline{u}_h)$ or $J_2(\underline{u}) - J_2(\underline{u}_h)$ and the ratio of successive error estimates. The asymptotic exactness value is as given in (8.2.22).

Tab. 8.2.15: This is for the functional J_1 with $b = 1$.

ne	Exact error	Estimates with quadratics	Asymptotic exactness value	Ratios
8	-4.28303e-006	-4.25881e-006	-5.655006e-003	
16	-2.60119e-007	-2.59627e-007	-1.891437e-003	16.403571
32	-1.61650e-008	-1.61568e-008	-5.048899e-004	16.069209
64	-1.00900e-009	-1.00887e-009	-1.282060e-004	16.014749

Tab. 8.2.16: This is for the functional J_1 with $b = 1/8$.

ne	Exact error	Estimates with quadratics	Asymptotic exactness value	
8	-1.35479e-007	-1.35194e-007	-2.106686e-003	
16	-8.93042e-009	-8.92635e-009	-4.553001e-004	15.145496
32	-5.67808e-010	-5.67744e-010	-1.118041e-004	15.722491
64	-3.56573e-011	-3.56563e-011	-2.781270e-005	15.922684

Tab. 8.2.17: This is for the functional J_2.

ne	Exact error	Estimates with quadratics	Asymptotic exactness value	Ratios
8	-4.06492e-004	-4.04811e-004	-4.134425e-003	
16	-2.58337e-005	-2.58068e-005	-1.040423e-003	15.686214
32	-1.62140e-006	-1.62098e-006	-2.606449e-004	15.920492
64	-1.01445e-007	-1.01439e-007	-6.529326e-005	15.979849

All these tables show that we get asymptotic exactness when the estimate is $-a(\underline{u}_h; \underline{\psi}_h^m)$. Also,similar with the previous tables, when we double the number of elements we decrease the error by about $2^4 = 16$.

Uniform refinement when the degree $p = 3$

For the deformation shown in figure 8.1(a) the approximation when $p = 3$ is very accurate with a small number of elements and we only give one table which supports the high accuracy and rapid convergence. The error estimate is $-a(\underline{u}_h; \underline{\psi}_h^m)$. The asymptotic exactness value is as given in (8.2.22).

Tab. 8.2.18: The exact error column means values $J_1(\underline{u}) - J_1(\underline{u}_h)$ in the case $b = 1/8$. and the column ratio has the ratio of successive error estimates.

ne	Exact error	Estimates with cubics	Asymptotic exactness value	Ratios
8	1.70229e-012	1.70829e-012	3.526218e-003	
16	2.73063e-014	2.73211e-014	5.430333e-004	62.526400

From the above table, we can see that when we double the number of elements we decrease the error by about $2^6 = 64$.

Therefore all the tables, where we use uniform refinement, confirm the theory about the estimate of the error which is about $O(h^{2p})$, with p being the degree of the polynomials in space.

Adaptive refinement when the degree $p = 1$

We now repeat the same set of problems considered when $p = 1$ and uniform refinement was done with instead the adaptive refinement procedure described in section 8.2.2. In all the cases we take *tol*$= 10^{-7}$ as the accuracy we wish to obtain.

The next two tables are for the functional J_1 when $b = 1$ and the deformation shown in figure 8.1(a). In both cases the exact error column means values $J_1(\underline{u}) - J_1(\underline{u}_h)$. The asymptotic exactness value is as given in (8.2.22).

Tab. 8.2.19: The error estimate is $-a(\underline{u}_h; \underline{\psi}_h^m)$.

ne	Exact error	Estimates with linears	Asymptotic exactness value
10	-4.92587e-005	-4.92492e-005	-1.932016e-004
144	-2.08504e-007	-2.08504e-007	-1.767264e-006
280	-5.58481e-008	-5.58481e-008	-7.310712e-007

Tab. 8.2.20: The error estimate is $-a(\underline{u}_h; \underline{\psi}_h)$.

ne	Exact error	Estimates with linears	Asymptotic exactness value
10	-4.92587e-005	-1.64796e-005	-6.654490e-001
115	-3.24680e-007	-1.03432e-007	-6.814335e-001
161	-1.86455e-007	-6.00604e-008	-6.778818e-001

The next two tables are for the functional J_1 when $b = 1/8$ and the deformation shown in figure 8.1(a). In both cases the exact error column means values $J_1(\underline{u}) - J_1(\underline{u}_h)$. The asymptotic exactness value is as given in (8.2.22).

Tab. 8.2.21: The error estimate is $-a(\underline{u}_h; \underline{\psi}_h^m)$.

ne	Exact error	Estimates with linears	Asymptotic exactness value
8	6.85176e-006	6.89246e-006	5.939109e-003
78	5.61210e-008	5.61242e-008	5.689184e-005

Tab. 8.2.22: The error estimate is $-a(\underline{u}_h; \underline{\psi}_h)$.

ne	Exact error	Estimates with linears	Asymptotic exactness value
8	6.85176e-006	7.65920e-006	1.178444e-001
77	4.43887e-008	5.98690e-008	3.487440e-001

The next two tables are for the functional J_2 and the deformation shown in figure 8.1(a). In both cases the exact error column means values $J_2(\underline{u}) - J_2(\underline{u}_h)$. The asymptotic exactness value is as given in (8.2.22).

Tab. 8.2.23: The error estimate is $-a(\underline{u}_h; \underline{\psi}_h^m)$.

ne	Exact error	Estimates with linears	Asymptotic exactness value
10	-3.25065e-005	-3.25064e-005	-3.451943e-006
140	-1.47692e-007	-1.47692e-007	-2.334666e-007
253	-4.83168e-008	-4.83168e-008	-9.333399e-008

Tab. 8.2.24: The error estimate is $-a(\underline{u}_h; \underline{\psi}_h)$.

ne	Exact error	Estimates with linears	Asymptotic exactness value
10	-3.25065e-005	-6.50109e-005	9.999353e-001
149	-1.34027e-007	-2.68055e-007	9.999993e-001
325	-2.78476e-008	-5.56951e-008	9.999997e-001

The results show that we can reach the required accuracy with 1 or 2 refinement steps when the error estimate is $-a(\underline{u}_h; \underline{\psi}_h^m)$. When we just use the error estimate is $-a(\underline{u}_h; \underline{\psi}_h)$ this does help to move towards the desired accuracy but we cannot guarantee that the result to all the digits given when we are underestimating the actual error.

Adaptive refinement when the degree $p = 2$

We now repeat the same same set of problems considered when $p = 2$ and uniform refinement was done with instead the adaptive refinement procedure described in section 8.2.2. In all the cases we take *tol*$= 10^{-10}$ as the accuracy we wish to obtain.

The next two tables are for the functional J_1 when $b = 1$ and the deformation shown in figure 8.1(a). In both cases the exact error column means values $J_1(\underline{u}) - J_1(\underline{u}_h)$. The asymptotic exactness value is as given in (8.2.22).

Tab. 8.2.25: The error estimate is $-a(\underline{u}_h; \underline{\psi}_h^m)$.

ne	Exact error	Estimates with quadratics	Asymptotic exactness value
10	-2.03137e-008	-2.03135e-008	-9.465689e-006
42	-5.15373e-011	-5.15371e-011	-5.383017e-006

Tab. 8.2.26: The error estimate is $-a(\underline{u}_h; \underline{\psi}_h)$.

ne	Exact error	Estimates with quadratics	Asymptotic exactness value
10	-2.03137e-008	-2.27888e-008	1.218432e-001
42	-5.15373e-011	-5.84564e-011	1.342536e-001

The next two tables are for the functional J_1 when $b = 1/8$ and the deformation shown in figure 8.1(a). In both cases the exact error column means values $J_1(\underline{u}) - J_1(\underline{u}_h)$. The asymptotic exactness value is as given in (8.2.22).

Tab. 8.2.27: The error estimate is $-a(\underline{u}_h; \underline{\psi}_h^m)$.

ne	Exact error	Estimates with quadratics	Asymptotic exactness value
8	-8.06079e-008	-8.05900e-008	-2.216413e-004
37	-5.85869e-011	-5.85871e-011	3.854325e-006

Tab. 8.2.28: The error estimate is $-a(\underline{u}_h; \underline{\psi}_h)$.

ne	Exact error	Estimates with quadratics	Asymptotic exactness value
8	-8.06079e-008	-8.08974e-008	3.591009e-003
37	-5.85869e-011	-5.99856e-011	2.387330e-002

The next two tables are for the functional J_2 and the deformation shown in figure 8.1(a). In both cases the exact error column means values $J_2(\underline{u}) - J_2(\underline{u}_h)$. The asymptotic exactness value is as given in (8.2.22).

Tab. 8.2.29: The error estimate is $-a(\underline{u}_h; \underline{\psi}_h^m)$.

ne	Exact error	Estimates with quadratics	Asymptotic exactness value
10	-4.01198e-009	-4.01201e-009	9.311140e-006
29	-4.97193e-011	-4.97193e-011	1.037262e-006

Tab. 8.2.30: The error estimate is $-a(\underline{u}_h; \underline{\psi}_h)$.

ne	Exact error	Estimates with quadratics	Asymptotic exactness value
10	-4.01198e-009	-8.02435e-009	1.000100e+000
34	-2.52926e-011	-5.05854e-011	1.000010e+000

The next three tables are for the deformation shown in figure 8.1(b) and in all cases the error estimate is $-a(\underline{u}_h; \underline{\psi}_h^m)$. In all cases the exact error column means values $J_1(\underline{u}) - J_1(\underline{u}_h)$ or $J_2(\underline{u}) - J_2(\underline{u}_h)$. The asymptotic exactness value is as given in (8.2.22).

Tab. 8.2.31: This is for the functional J_1 with $b = 1$.

ne	Exact error	Estimates with quadratics	Asymptotic exactness value
8	-4.28303e-006	-4.25881e-006	-5.655006e-003
128	-6.30422e-011	-6.30403e-011	-3.036950e-005
418	-4.04239e-013	-4.04357e-013	2.909084e-004

Tab. 8.2.32: This is for the functional J_1 with $b = 1/8$.

ne	Exact error	Estimates with quadratics	Asymptotic exactness value
8	-1.35479e-007	-1.35194e-007	-2.106686e-003
126	-2.28389e-012	-2.28387e-012	-6.981137e-006
223	-3.01674e-013	-3.01673e-013	-5.942295e-006

Tab. 8.2.33: This is for the functional J_2.

ne	Exact error	Estimates with quadratics	Asymptotic exactness value
8	-4.06492e-004	-4.04811e-004	-4.134425e-003
128	-6.34203e-009	-6.34193e-009	-1.566206e-005
1109	-7.38520e-013	-7.24830e-013	-1.853723e-002

As in the case $p = 1$, all the results show that we can reach the required accuracy with 1 or 2 refinement steps when the error estimate is $-a(\underline{u}_h; \underline{\psi}_h^m)$.

Adaptive refinement when the degree $p = 3$

The next three tables are for the deformation shown in figure 8.1(a). In all cases the exact error column means values $J_1(\underline{u}) - J_1(\underline{u}_h)$ or $J_2(\underline{u}) - J_2(\underline{u}_h)$. In all cases the error estimate is $-a(\underline{u}_h; \underline{\psi}_h^m)$. The asymptotic exactness value is as given in (8.2.22). For the adaptive refinement algorithm *tol*= 10^{-14}.

Tab. 8.2.34: This is for the functional J_1 with $b = 1$.

ne	Exact error	Estimates with cubics	Asymptotic exactness value
10	-2.58460e-013	-2.58485e-013	9.797869e-005
18	5.55112e-016	5.51636e-016	-6.260532e-003

Tab. 8.2.35: This is for the functional J_1 with $b = 1$.

ne	Exact error	Estimates with cubics	Asymptotic exactness value
8	1.70229e-012	1.70829e-012	3.526218e-003
21	2.18402e-015	2.18819e-015	1.911335e-003

Tab. 8.2.36: This is for the functional J_2.

ne	Exact error	Estimates with cubics	Asymptotic exactness value
10	-1.65597e-014	-1.65582e-014	-9.040961e-005
15	-2.81025e-015	-2.83074e-015	7.290249e-003

In the cases here we can reach the required accuracy with 1 refinement step.

8.3 The dynamic problem

8.3.1 Introduction

In this section we present results using the numerical methods described in chapter 6 for the dynamic problem, i.e. when the full equations of motion are involved. The faster the pressure is applied the greater the difference we can expect between the quasi-static solution and the dynamic solution for the same pressure loading and we consider this first before quantities of interest are considered. Let

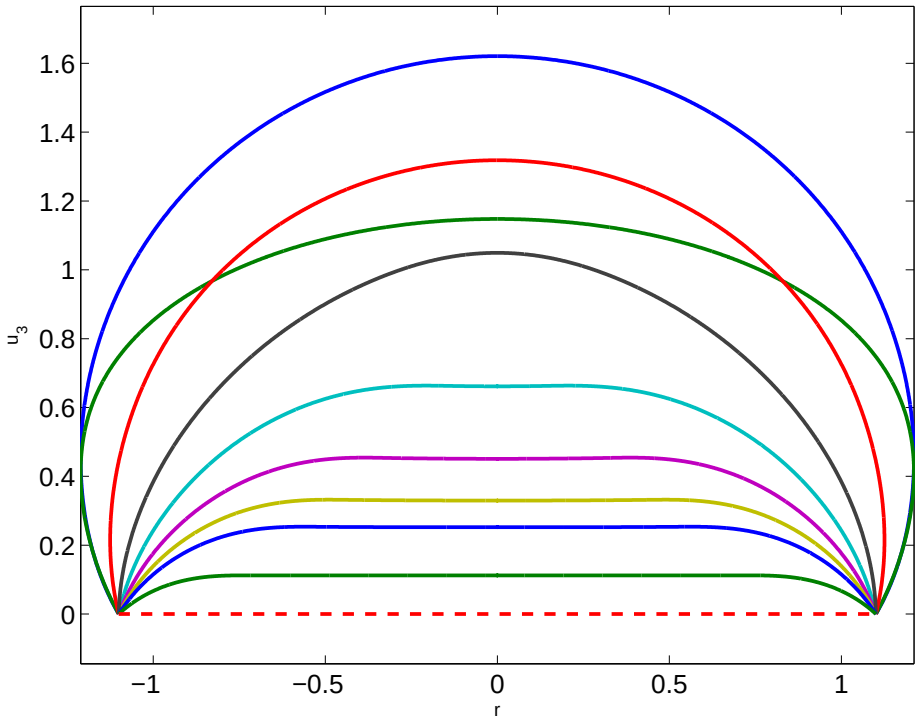
$$Q = [0,1] \times [0,T] = \{(r,t) : \ 0 \leq r \leq 1, \ 0 \leq t \leq T\}$$

denote the space-time domain and assume that the applied pressure varies with t according to

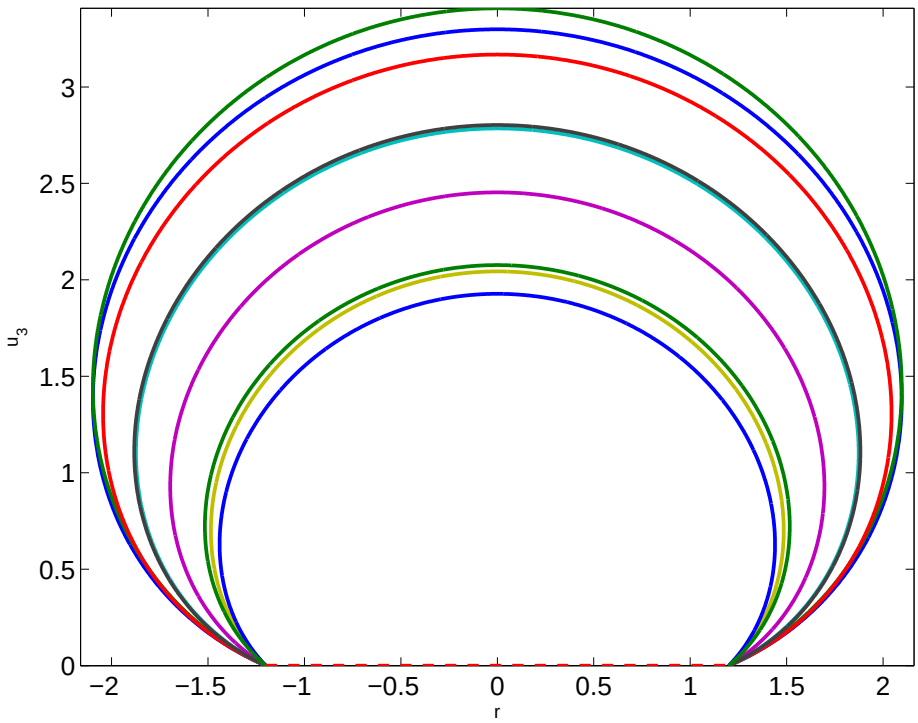
$$P(t) = \gamma t, \quad 0 \leq t \leq T, \tag{8.3.1}$$

where γ is a constant. To compare the solutions at the same pressure for different rates we take different values of γ and T as follows. For the Jones Treloar model the values are such that $P(T) = 1$ and the profiles at time $t = T$ in each case are shown in figure 8.2(a). For the Mooney Rivlin model the values are such that $P(T) = 3$ and the profiles at time $t = T$ in each case are shown in figure 8.2(b). In each of the figures the outermost curve corresponds to the smallest value of γ. In the Mooney Rivlin case the outermost curve is close to the outermost curve for the quasi-static case shown in figure 8.1(b) but apart from this case there is a noticeable difference between most of the profiles.

As well as considering the profiles, which are at the fixed time $t = T$, it is also interesting to show how some quantities vary with time t for a fixed value of r. In the case of the profile corresponding to $\gamma = 0.25$ in figure 8.2(a) we show in figures 8.3(a)–8.3(d) graphs of $u_3(r,t)$, $0 \leq t \leq T$ for $r = 0$, 0.25, 0.5 and 0.75. Similarly, in the case of the profile corresponding to $\gamma = 1$ in figure 8.2(a) we show in figures 8.4(a)–8.4(d) graphs of $u_3(r,t)$, $0 \leq t \leq T$ for $r = 0$, 0.25, 0.5 and 0.75. As the figures in 8.3(a)–8.3(d) show, the vertical displacement does not always monotonically increase as the pressure increases as it mostly does with the quasi-static inflations.

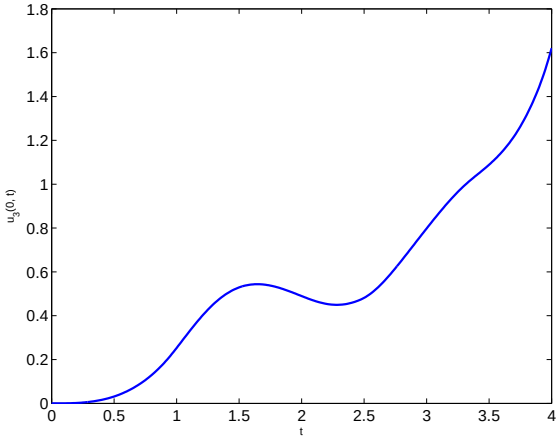


(a) Jones Treloar material with $P(T) = 1$ and for $\gamma = 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2, 3$

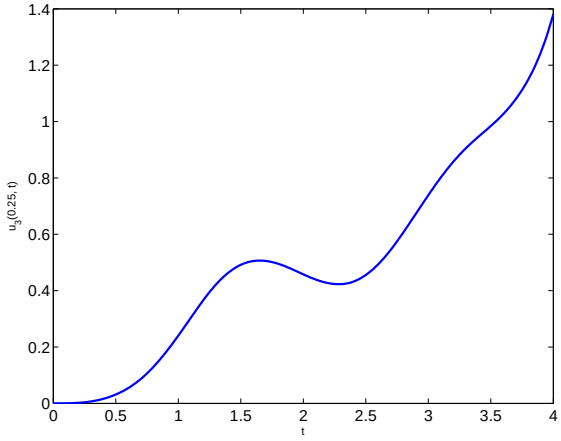


(b) Mooney Rivlin material with $P(T) = 3$ and for $\gamma = 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2, 3$

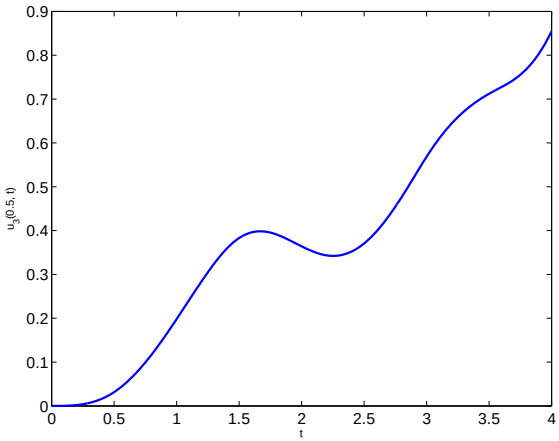
Fig. 8.2: In each figure all the profiles are at the same pressure but correspond to different rates at which it has been applied.



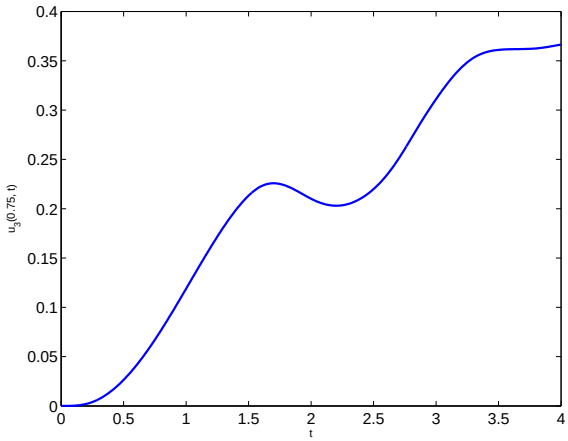
(a) A graph of $u_3(0,t)$, $0 \leq t \leq T$.



(b) A graph of $u_3(0.25,t)$, $0 \leq t \leq T$.

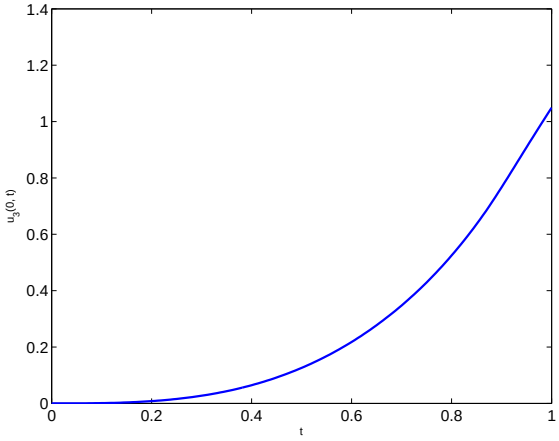


(c) A graph of $u_3(0.5,t)$, $0 \leq t \leq T$.

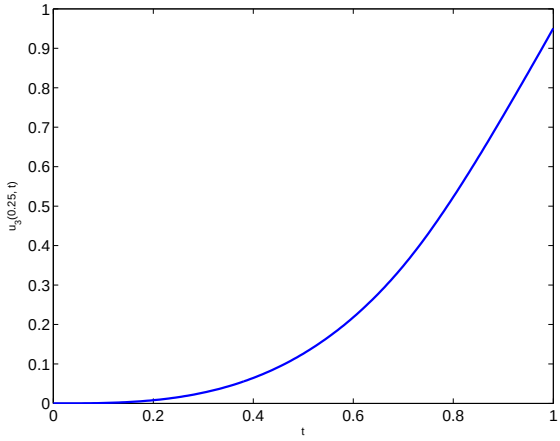


(d) A graph of $u_3(0.75,t)$, $0 \leq t \leq T$.

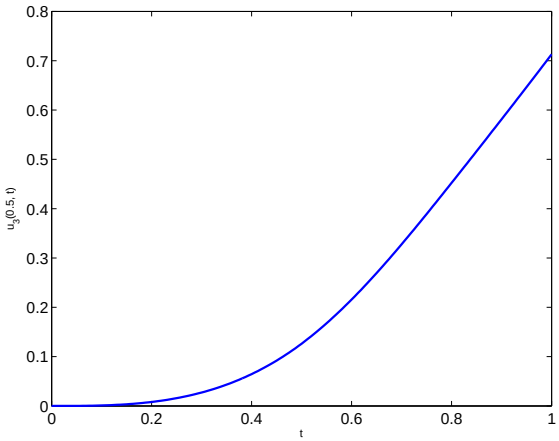
Fig. 8.3: In each figure all the paths are for fixed values of r for the case $\gamma = 0.25$ and for the Jones Treloar material.



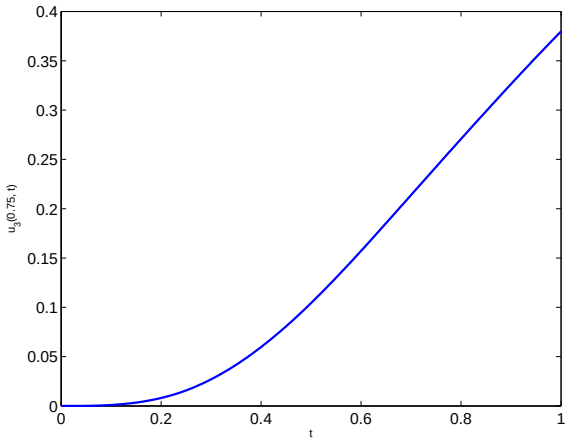
(a) A graph of $u_3(0,t)$, $0 \leq t \leq T$.



(b) A graph of $u_3(0.25,t)$, $0 \leq t \leq T$.



(c) A graph of $u_3(0.5,t)$, $0 \leq t \leq T$.



(d) A graph of $u_3(0.75,t)$, $0 \leq t \leq T$.

Fig. 8.4: In each figure all the paths are for fixed values of r for the case $\gamma = 1$ and for the Jones Treloar material.

In this section we consider two quantity of interest functionals which are as follows.

$$J_3\left(\begin{pmatrix}\underline{u}\\ \underline{v}\end{pmatrix}\right) = \int_0^b u_3(r,T)rdr, \quad b = 0.1 \tag{8.3.2}$$

and

$$J_4\left(\begin{pmatrix}\underline{u}\\ \underline{v}\end{pmatrix}\right) = \frac{2}{b^2}\frac{1}{0.01T}\int_{0.99T}^{T}\int_0^b \lambda(r,t)rdrdt \quad b = 0.1. \tag{8.3.3}$$

J_3 involves the vertical displacement u_3 near the pole at the final time and if we divide by $b^2/2$ then we get the average of u_3 near the pole at time $t = T$. We cannot at present cope with the thickness just at the final time but we can handle the expression for J_4 which involves an average of the thickness stretch ratio near the final time and near the pole. For the Gâteaux derivative of J_3 we have

$$J_3'\left(\begin{pmatrix}\underline{u}\\ \underline{v}\end{pmatrix};\begin{pmatrix}\underline{\alpha}\\ \underline{\beta}\end{pmatrix}\right) = \int_0^b \underline{\alpha}(r,T)\cdot\underline{J}_\alpha(r,T)rdr = \int_0^b \alpha_3(r,T)rdr. \tag{8.3.4}$$

In the case of J_4 we have, similar to (8.2.6) in the quasi-static case,

$$\begin{aligned} &J_4'\left(\begin{pmatrix}\underline{u}\\ \underline{v}\end{pmatrix};\begin{pmatrix}\underline{\alpha}\\ \underline{\beta}\end{pmatrix}\right)\\ &\quad = \frac{2}{b^2}\frac{1}{0.01T}\int_{0.99T}^{T}\int_0^b -(\lambda_1\lambda_2)^{-2}\left(\lambda_1\frac{\alpha_1}{r}+\frac{\lambda_2}{\lambda_1}((1+u_1)\alpha_1'+u_3'\alpha_3')\right)rdrdt. \end{aligned} \tag{8.3.5}$$

Now to simplify the notation a little when both $\underline{u}$ and $\underline{v}$ are unknowns we write the following.

$$\underline{U} = \begin{pmatrix}\underline{u}\\ \underline{v}\end{pmatrix}, \quad \underline{U}_h = \begin{pmatrix}\underline{u}_h\\ \underline{v}_h\end{pmatrix}, \quad \underline{U}_h^b = \begin{pmatrix}\underline{u}_h^b\\ \underline{v}_h^b\end{pmatrix}, \quad \underline{U}_h^m = \begin{pmatrix}\underline{u}_h^m\\ \underline{v}_h^m\end{pmatrix} \quad \text{and} \quad \underline{z}_h^m = \begin{pmatrix}\underline{\psi}_h^m\\ \underline{\theta}_h^m\end{pmatrix}.$$

In the examples considered here, we have the Jones Treloar form for W, $h_0 = 1$, the density $\rho = 0.2$, we start with a prestretch such that $u_1(1,t) = 0.1$ and the pressure rate γ and the final time T are such that $P(T) = 0.3$. For one of the ways of approximating J_3 and J_4 we consider a range of the values γ and later we restrict to the case $\gamma = 0.1$. The aims of the computations are to get accurate approximations, to test if we can accurately estimate the error in an approximation and to help to determine what we need to do to improve the accuracy. In this investigation we consider both the "basic scheme in time" which uses degree $n = 1$ polynomials in t for $\underline{u}_h(r,t)$ and $\underline{v}_h(r,t)$ in a time interval $t_{j-1} < t < t_j$ and we consider the higher order in time scheme described in the latter part

of chapter 6 which use larger values of the degree n. For the larger space used to obtain the dual solution $\underline{\psi}_h^m(r,t)$ and $\underline{\theta}_h^m(r,t)$ we take the following.

(i) When we use the basic scheme for V_h with degree p polynomials on each of ne elements in space and with nt time steps using degree $n = 1$ polynomials in each time interval the larger space $\bar{V}_h$ uses degree $p+1$ polynomials on each of the ne elements and it also uses degree $n = 1$ polynomials in each time interval but the number of time intervals is doubled.

(ii) When for V_h we use degree $p \geq 1$ polynomials on each of ne elements in space and we use degree $n > 1$ polynomials in time on each of the nt time intervals the larger space $\bar{V}_h$ uses degree $p+1$ polynomials on each of the ne elements and it uses degree $n+1$ polynomials on each of the nt time intervals.

The larger space $\bar{V}_h$ is also the space used when we want to get a better solution $\underline{u}_h^b$ for the displacement and $\underline{v}_h^b$ for the velocity in order to define the mid-point values

$$\underline{u}_h^m = \frac{1}{2}\left(\underline{u}_h + \underline{u}_h^b\right), \quad \underline{v}_h^m = \frac{1}{2}\left(\underline{v}_h + \underline{v}_h^b\right) \tag{8.3.6}$$

to use as data in the dual problem. When a better approximation is obtained we can compare the dual solution estimate

$$J\left(\underline{U}\right) - J\left(\underline{U}_h\right) \approx F(\underline{z}_h^m) - A\left(\underline{U}_h; \underline{z}_h^m\right) \tag{8.3.7}$$

with the estimate

$$J\left(\underline{U}\right) - J\left(\underline{U}_h\right) \approx J\left(\underline{U}_h^b\right) - J\left(\underline{U}_h\right). \tag{8.3.8}$$

When a possibly better approximation $\underline{u}_h^b$ and $\underline{v}_h^b$ is available it is likely that the quantity of interest value $J\left(\underline{U}_h^b\right)$ is the best estimate that we have of the quantity of interest but we have no information about its accuracy and we do not have too much information to help determine what should be done in terms of refining to get a more accurate approximation. Also, as some examples show, the value is only much more accurate when we are in the asymptotic convergence range and in practice this has to be determined during the computation.

8.3.2 The basic scheme – results for different pressure rates

We investigate here the effect the rate at which the pressure is applied has on the quantity of interest values and how well we estimate the error in the quantity of interest in each case when we use the basic scheme in time. In each case we do this when at the the final time $P(T) = 0.3$ with the 4 rates $\gamma = 0.1, 0.15, 0.2$ and 0.5. The numerical approximation is obtained in each case using *ne=10* quadratic elements in space and *nt=100* equal time steps. The deformation is never too large in any of these cases and to be specific we give the vertical displacement $u_3(0,T)$ that we obtain in the table 8.3.37. The values for J_3 and J_4 given to compute the column indicated as "exact error" were obtained with a higher order scheme. In each case the error estimate obtained using the solution to a dual problem always has a similar magnitude and sign as the exact error.

Tab. 8.3.37: The vertical displacement $u_3(0,T)$ when $P(T) = \gamma T = 0.3$ for different values of the rate γ.

γ	$u_3(0,T)$
0.1	0.2923733
0.15	0.3705022
0.2	0.4431038
0.5	0.1089220

Tab. 8.3.38: The exact value and the estimate of J_3, the exact error and the error estimates when *ne=10*, $p = 2$ and *nt=100* with different inflation rates γ

γ	J_3	Estimates	Exact error	Error Estimates
0.10	1.455296e-03	1.453634e-03	1.661909e-06	1.237341e-06
0.15	1.842955e-03	1.843510e-03	-5.547729e-07	-4.147575e-07
0.20	2.202974e-03	2.202998e-03	-2.385906e-08	-1.547300e-08
0.50	5.446210e-04	5.446632e-04	-4.222632e-08	-3.483284e-08

Tab. 8.3.39: The exact value and the estimate of J_4, the exact error and the error estimates when *ne=10*, $p = 2$ and *nt=100* with different inflation rates γ

γ	J_4	Estimates	Exact error	Error Estimates
0.10	7.511797e-01	7.511483e-01	3.137137e-05	1.819369e-05
0.15	7.069278e-01	7.068821e-01	4.574654e-05	3.098082e-05
0.20	6.851688e-01	6.851881e-01	-1.933876e-05	-1.773006e-05
0.50	8.211742e-01	8.211750e-01	-8.087841e-07	-1.928018e-06

8.3.3 The basic scheme – results with increasing ne *and* nt

The tables 8.3.38 and 8.3.39 show that we get estimates of the error which we might regard as acceptable although they are never very close to the exact error. The investigation here is to determine, for a fixed value of the pressure rate γ, how the accuracy improves as we uniformly refine in space and/or time. Eventually, to improve the accuracy it is necessary to refine in both space and time although it may be wasteful to do this from the start when the error may be mostly due to just one of the space discretization error and the time discretization error. The rate $\gamma = 0.1$ is used in all the result in this section and to be consistent with the tables 8.3.38 and 8.3.39 we take quadratics (i.e. $p = 2$) and we start with *ne=10* and *nt=100* in most cases. The results that we get from first doubling *nt* and then doubling *ne* are shown in tables 8.3.40 and 8.3.41.

Tab. 8.3.40: The estimates of J_3 when $p = 2$ and $n = 1$, the exact error and the error estimates where we perform one uniform refinement in space and one refinement step in time. The exact QoI with the inflation rate of $\gamma = 0.1$ is J_3 =1.45529590888099e-003.

ne	*nt*	Estimates of J_3	Exact error	Error Estimates
10	100	1.453634e-03	1.661909e-06	1.237341e-06
10	200	1.454858e-03	4.379089e-07	3.313185e-07
20	200	1.454864e-03	4.319089e-07	3.245756e-07

Tab. 8.3.41: The estimates of J_4 when $p = 2$ and $n = 1$, the exact error and the error estimates where we perform one uniform refinement in space and one refinement step in time. The exact QoI with the inflation rate of $\gamma = 0.1$ is J_4 =7.51179671367555e-001.

ne	*nt*	Estimates of J_4	Exact error	Error Estimates
10	100	7.511483e-01	3.137137e-05	1.819369e-05
10	200	7.511733e-01	6.371368e-06	3.637537e-06
20	200	7.511676e-01	1.207137e-05	8.662038e-06

Although the tables 8.3.40 and 8.3.41 do not contain many numbers there are enough to strongly suggest that to improve the accuracy of J_3 we need more time steps and the error at this stage seems to be almost entirely due to the time discretization error. The results for J_4 are less clear. To get a better idea of how things change with *ne* and *nt* we show in tables 8.3.42 and 8.3.43 more combinations of *ne* and *nt* with in each case just the error estimate being shown. In the case of J_3 the increase in accuracy is entirely due to just taking more time steps and we have not detected a stage when we need to increase *ne* from 10. In the case of J_4 we need to increase *ne* from 10 to 20 but it is not necessary to use *ne*=40 for the number of time steps given. As a consequence of these observations we show in tables 8.3.44 and 8.3.45 the ratio of successive error estimates when we just successively double *nt* with *ne*=10 for J_3 and *ne*=20 for J_4 and in all cases these are about 4. For the examples considered, we need a large number of time steps compared

with the number of elements in space when $p = 2$ and as doubling nt only reduces the error by a factor of about 4, high accuracy with the basic scheme needs very large values of nt. It is as a consequence of this that the higher order in time schemes were described in chapter 6 and results for the higher order schemes are presented in subsequent sections.

Tab. 8.3.42: The estimates of J_3, with a range of values for the number of time steps nt and a range of values for the number of elements ne when $p = 2$ and the inflation rate $\gamma = 0.1$

ne \ nt	50	100	200	400	800
10	8.058503e-006	1.237341e-006	3.313185e-007	7.563159e-008	1.691794e-008
20	8.090020e-006	1.250205e-006	3.245756e-007	8.013757e-008	2.067992e-008
40	8.093398e-006	1.257161e-006	3.163513e-007	8.026591e-008	2.009467e-008

Tab. 8.3.43: The estimates of J_4, with a range of values for the number of time steps nt and a range of values for the number of elements ne when $p = 2$ and the inflation rate $\gamma = 0.1$

ne \ nt	100	200	400	800
10	1.819369e-005	3.637537e-006	-1.320211e-005	-1.813363e-005
20	2.573690e-005	8.662038e-006	1.915007e-006	4.808166e-007
40	2.690115e-005	9.484209e-006	2.408255e-006	5.785791e-007

Tab. 8.3.44: The estimates of J_3 for number of elements of $ne = 10$, with different number of time steps nt with the corresponding ratios.

nt	50	100	200	400	800
est	8.058503e-006	1.237341e-006	3.313185e-007	7.563159e-008	1.691794e-008
ratio		6.5128	3.7346	4.3807	4.4705

Tab. 8.3.45: The estimates of J_4 for number of elements of $ne = 20$, with different number of time steps nt with the corresponding ratios.

nt	100	200	400	800
est	2.573690e-005	8.662038e-006	1.915007e-006	4.808166e-007
ratio		2.971229	4.523241	3.982822

8.3.4 The higher order scheme – experiments with different values of ne, nt, *p and n*

A reasonable deduction from the results with the basic scheme is that the error in approximating the quantity of interests considered in this section is influenced significantly by how well or otherwise we approximate in time and with the basic scheme the accuracy in time is low. With the basic scheme in the case of J_3, $p = 2$ and ne=10 the successive doubling of nt up to 800 was not sufficient to reach the stage that we need to also increase ne to improve the accuracy. Similarly, in the case of J_4, $p = 2$ and ne=20 the successive doubling of nt up to 800 was also not sufficient to reach the stage that we need to also increase ne to improve the accuracy. We consider both cases again here with again $p = 2$ for the degree of polynomials in r on each element but we now present results with $n = 2$, $n = 3$ and $n = 4$ to attempt to get to the stage when we need to increase the number of elements ne to improve the accuracy. This is done for J_3 in tables 8.3.46 and 8.3.47 in respectively the cases ne=10 and ne=20 and for J_4 in tables 8.3.48 and 8.3.49 in respectively the cases ne=20 and ne=40. In the case of J_3, ne=10 and $n = 2$ there is no longer need to increase nt after nt=200, it is nt=100 when $n = 3$ and nt=50 is already large enough when $n = 4$. The results for larger values of nt are unnecessary, as they do not lead to any decrease in the error, although they do confirm that the computations are likely to be correct. When we increase ne to 20 the smallest values of nt such that the ratio is close to 1 are the same or a bit higher than they were with ne=10. It is a similar pattern in the case of J_4 where it is nt=400, 200, and 100 for $n = 2$, 3 and 4 respectively when ne=20. When ne=40 these values of nt are instead 800, 400 and 100 for $n = 2$, 3 and 4 respectively.

As a final remark here, to have some confidence in the accuracy we want the value in the 'ratio' column in the tables to be close to 1 and to avoid unnecessary computation we should stop increasing nt at this stage. From the results it is also worth noting that in most cases when the ratio value is not close to 1 both the dual solution estimate and the estimate $J(\underline{U}_h^b) - J(\underline{U}_h)$ are far from the true error in that they are far from the value obtained when nt is larger.

Tab. 8.3.46: The QoI J_3, ne=10 elements, $p = 2$ for the degree of the polynomials in space, values of the dual solution estimate, the estimate $J_3(\underline{U}_h^b) - J_3(\underline{U}_h)$ and the ratio of these 2 estimates.

nt	n	Dual sol. estimate	$J_3(\underline{U}_h^b) - J_3(\underline{U}_h)$	Ratio
50	2	3.685873e-08	-6.311311e-10	-58.401060
100	2	-3.735936e-09	-4.797271e-09	0.778763
200	2	-2.258335e-09	-2.346515e-09	0.962421
400	2	-2.239576e-09	-2.245107e-09	0.997536
800	2	-2.239162e-09	-2.239516e-09	0.999842
50	3	-3.333265e-09	-4.365725e-09	0.763508
100	3	-2.206283e-09	-2.223503e-09	0.992256
200	3	-2.238295e-09	-2.238876e-09	0.999741
400	3	-2.239123e-09	-2.239139e-09	0.999993
800	3	-2.239136e-09	-2.239143e-09	0.999997
50	4	-2.185222e-09	-2.129653e-09	1.026093
100	4	-2.237262e-09	-2.238104e-09	0.999624
200	4	-2.239130e-09	-2.239139e-09	0.999996
400	4	-2.239136e-09	-2.239143e-09	0.999997
800	4	-2.239136e-09	-2.239143e-09	0.999997

Tab. 8.3.47: The QoI J_3, ne=20 elements, $p = 2$ for the degree of the polynomials in space, values of the dual solution estimate, the estimate $J_3(\underline{U}_h^b) - J_3(\underline{U}_h)$ and the ratio of these 2 estimates.

nt	n	Dual sol. estimate	$J_3(\underline{U}_h^b) - J_3(\underline{U}_h)$	Ratio
50	2	1.421935e-07	1.309190e-08	10.861180
100	2	1.659497e-08	-7.606752e-10	-21.816110
200	2	1.252540e-09	1.135255e-09	1.103311
400	2	9.859472e-10	9.765729e-10	1.009599
800	2	9.672640e-10	9.665601e-10	1.000728
50	3	-1.732789e-09	-2.286074e-10	7.579757
100	3	1.081486e-09	1.181619e-09	0.915258
200	3	9.692134e-10	9.662741e-10	1.003042
400	3	9.659856e-10	9.659745e-10	1.000011
800	3	9.659791e-10	9.659784e-10	1.000001
50	4	1.047201e-09	9.090907e-10	1.151922
100	4	9.760658e-10	9.726916e-10	1.003469
200	4	9.659004e-10	9.659488e-10	0.999950
400	4	9.659787e-10	9.659782e-10	1.000000
800	4	9.659789e-10	9.659785e-10	1.000000

Tab. 8.3.48: The QoI J_4, ne=20 elements, $p = 2$ for the degree of the polynomials in space, values of the dual solution estimate, the estimate $J_4(\underline{U}_h^b) - J_4(\underline{U}_h)$ and the ratio of these 2 estimates.

nt	n	Dual sol. estimate	$J_4(\underline{U}_h^b) - J_4(\underline{U}_h)$	Ratio
100	2	-2.307075e-06	-2.527944e-07	9.126291
200	2	-3.731176e-07	4.262622e-07	-0.875324
400	2	6.240813e-07	6.212356e-07	1.004581
800	2	6.300560e-07	6.299383e-07	1.000187
100	3	3.020121e-07	5.442134e-07	0.554952
200	3	6.330231e-07	6.307151e-07	1.003659
400	3	6.304799e-07	6.304617e-07	1.000029
800	3	6.304522e-07	6.304522e-07	1.000000
100	4	6.355211e-07	6.317728e-07	1.005933
200	4	6.309598e-07	6.304821e-07	1.000758
400	4	6.304517e-07	6.304522e-07	0.999999
800	4	6.304520e-07	6.304520e-07	1.000000

Tab. 8.3.49: The QoI J_4, ne=40 elements, $p = 2$ for the degree of the polynomials in space, values of the dual solution estimate, the estimate $J_4(\underline{U}_h^b) - J_4(\underline{U}_h)$ and the ratio of these 2 estimates.

nt	n	Dual sol. estimate	$J_4(\underline{U}_h^b) - J_4(\underline{U}_h)$	Ratio
100	2	-8.720207e-06	1.488634e-08	-585.786000
200	2	9.052163e-07	-2.084614e-08	-43.423680
400	2	7.427578e-08	-5.415909e-08	-1.371437
800	2	-6.087182e-08	-6.039886e-08	1.007831
100	3	-6.081393e-08	-2.729508e-08	2.228018
200	3	-3.785711e-08	-5.904958e-08	0.641107
400	3	-6.070952e-08	-6.079825e-08	0.998541
800	3	-6.081567e-08	-6.081789e-08	0.999964
100	4	-5.605110e-08	-5.694790e-08	0.984252
200	4	-6.074129e-08	-6.079084e-08	0.999185
400	4	-6.084581e-08	-6.081838e-08	1.000451
800	4	-6.081814e-08	-6.081819e-08	0.999999

8.3.5 The higher order scheme – attempts at refining to achieve a specified accuracy

When the much easier quasi-static problems were considered the theory and the examples showed that when $p = 2$ and we are in the asymptotic convergence range the doubling of the number of elements ne leads to a decrease in the error in the estimate of our quantities of interest by a factor of about $2^{2p} = 16$. We consider next a strategy to attempt to get a similar reduction in the error in this more difficult dynamic case from one or two refinement steps. From all the tables presented so far, this is only going to happen when the error due to the time discretization is sufficiently small and also the two error estimators only appear to be reliable when their ratio is close to 1. From the experiments with different values of n when $p = 2$ given in tables 8.3.46–8.3.49 we get ratios which are close to 1 quite quickly when $n \geq 3$. Based on these observations we choose to take n=3 when $p = 2$ and do the following.

Step 1: With an initial number of elements ne and time steps nt we compute all the quantities and compute the ratio of the error estimates. If the ratio is in the interval $(0.7, 1.3)$ then we go to step 3.

Step 2: We successively replace nt by $2nt$, we compute all the quantities and we compute the ratio of the error estimates. When the ratio is in the interval $(0.7, 1.3)$ then we go to step 3.

Step 3: If the last value of the ratio of the error estimates is in $(0.7, 1.3)$ then we replace ne by $2ne$, we replace nt by $2nt$ and we repeat all the computations.

Step 4: If the last value of the ratio of the error estimates is not in $(0.7, 1.3)$ then we just replace nt by $2nt$ and we repeat all the computations.

Step 5: Go to step 3 or stop the computations if a desired accuracy has been reached or we have reached the limit of the number of elements or the number of time steps we wish to use.

In the case of the functionals J_3 and J_4, $\gamma = 0.1$ and a maximum pressure of $P(T) = \gamma T = 0.3$ we show the results of this approach in tables 8.3.50 and 8.3.51. In each table there is just one stage when we just double nt in our attempt to get back into the asymptotic convergence range and hence the strategy works well.

As already mentioned the deformation when $\gamma = 0.1$ and $P(T) = 0.3$ is not too large with the vertical displacement $u_3(0, T)$ being just below 0.3 as given in table 8.3.37. We

now consider a much larger deformation when we replace $P(T) = 0.3$ by $P(T) = 0.95$ which gives $u_3(0,T) \approx 1.7$. The results for J_3 and J_4 are shown in tables 8.3.52 and 8.3.53. In both cases there are more steps when we just double nt but again this crude strategy works quite well.

Tab. 8.3.50: The error estimates for the functional J_3 when $\gamma = 0.1$, $P(T) = \gamma T = 0.3$, $p = 2$ for the degree of the polynomials in space and $n = 3$ for the degree of polynomials in time. The most accurate approximation of $J_3(\underline{U})$ is `1.45529590910553e-03`.

ne	nt	Dual sol. estimate	$J_3(\underline{U}_h^b) - J_3(\underline{U}_h)$	Ratio	Near 1
10	50	-3.333265e-09	-4.365725e-09	0.763508	Y
20	100	1.081486e-09	1.181619e-09	0.915258	Y
40	200	6.351612e-11	3.848201e-11	1.650541	N
40	400	4.055883e-11	4.047279e-11	1.002126	Y
80	800	4.992943e-13	5.022619e-13	0.994092	Y
160	1600	1.401808e-13	1.398077e-13	1.002669	Y

Tab. 8.3.51: The error estimates for the functional J_4 when $\gamma = 0.1$, $P(T) = \gamma T = 0.3$, $p = 2$ for the degree of the polynomials in space and $n = 3$ for the degree of polynomials in time. The most accurate approximation of $J_4(\underline{U})$ is `7.51179671377779e-01`.

ne	nt	Dual sol. estimate	$J_4(\underline{U}_h^b) - J_4(\underline{U}_h)$	Ratio	Near 1
20	100	3.020121e-07	5.442134e-07	0.554952	N
20	200	6.330231e-07	6.307151e-07	1.003659	Y
40	400	-6.070952e-08	-6.079825e-08	0.998541	Y
80	800	-1.817206e-09	-1.787994e-09	1.016338	Y
160	1600	-7.710433e-11	-9.710333e-11	0.794044	Y

Tab. 8.3.52: The error estimates for the functional J_3 when $\gamma = 0.1$, $P(T) = \gamma T = 0.95$, $p = 2$ for the degree of the polynomials in space and $n = 3$ for the degree of polynomials in time. The most accurate approximation of $J_3(\underline{U})$ is `8.39753468720017e-03`

ne	nt	Dual sol. estimate	$J_3(\underline{U}_h^b) - J_3(\underline{U}_h)$	Ratio	Near 1
10	50	-3.283564e-06	-3.644651e-07	9.009270	N
10	100	-1.950239e-07	-5.868776e-08	3.323075	N
10	200	4.095176e-09	-5.278092e-09	-0.775882	N
10	400	-6.983998e-09	-6.862918e-09	1.017643	Y
20	800	8.508914e-09	8.504579e-09	1.000510	Y
40	1600	2.357069e-10	2.356552e-10	1.000220	Y
80	3200	6.428341e-12	6.446147e-12	0.997238	Y

Tab. 8.3.53: The error estimates for the functional J_4 when $\gamma = 0.1$, $P(T) = \gamma T = 0.95$, $p = 2$ for the degree of the polynomials in space and $n = 3$ for the degree of polynomials in time. The most accurate approximation of $J_4(\underline{U})$ is `1.04836176193406e-01`

ne	nt	Dual sol. estimate	$J_4(\underline{U}_h^b) - J_4(\underline{U}_h)$	Ratio	Near 1
20	100	4.170084e-07	-5.943337e-07	-0.701640	N
20	200	-8.150731e-07	-1.075281e-06	0.758010	Y
40	400	-5.730808e-08	-9.865345e-08	0.580903	N
40	800	-1.038427e-07	-1.044459e-07	0.994225	Y
80	1600	-7.655671e-09	-7.314915e-09	1.046584	Y
160	3200	-5.000234e-10	-5.090594e-10	0.982249	Y

8.3.6 Concluding remarks about the results

From all the results presented we have demonstrated a way to get high accuracy and we have demonstrated that the error estimate that we get after first solving a dual problem is accurate once we get into the asymptotic convergence range. One of the most difficult aspects to overcome is to get to the stage when we are in the asymptotic convergence range and this has required a lot of effort into how we approximate in time as well as taking significantly more time steps nt compared with the number of space elements ne. We have not yet reached the stage of trying to refine in space and/or time in a non-uniform way based on considering the expression for $F(\underline{z}_h^m) - A(\underline{U}_h; \underline{z}_h^m)$ with integrals just over part of the space time domain although we need a problem where an adaptive refinement is significant saving compared to refining all elements and all time intervals. The last example in section 8.3.5 when $P(T) = 0.95$ is a possible candidate problem to consider as the inflation is more rapid near the final time $t = T$ than it is for smaller times.

We have not tried too many quantities of interest in the dynamic case yet and we are restricted to what we can handle when the quantity of interest only involves the final time. For example, it would be reasonable to want to consider the functional

$$J_5(\underline{U}) = \int_0^b \lambda(r, T)\, r dr \tag{8.3.9}$$

instead of the functional J_4 which involves the thickness near the final time. The difficulty with J_5 is that the Gâteaux derivative is of the form

$$J_5'(\underline{U}) = \int_0^b \left(J_{\alpha_1}\alpha_1 + J_{\alpha_1'}\alpha_1' + J_{\alpha_3'}\alpha_3' \right)\, r dr \tag{8.3.10}$$

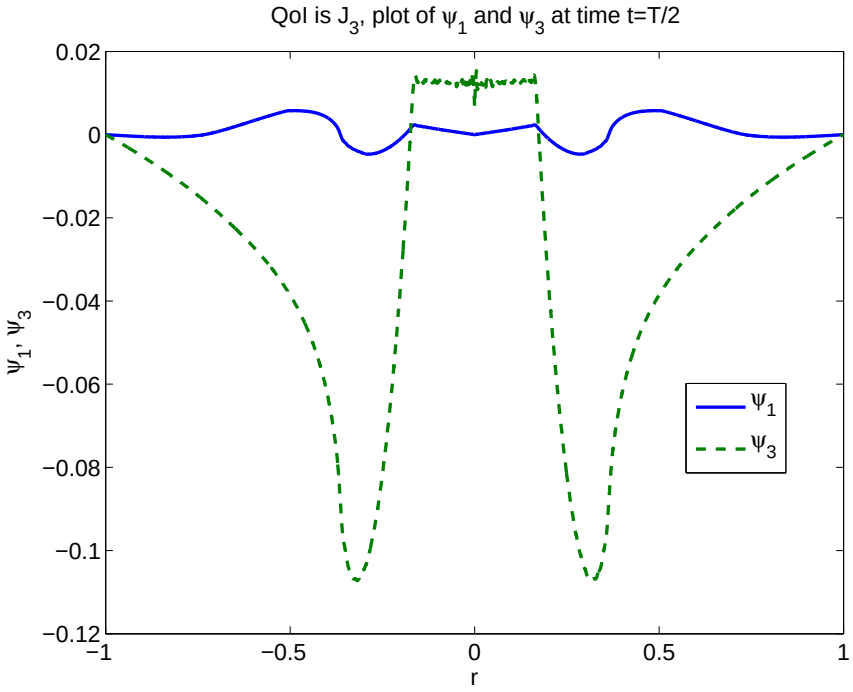
and the equation to determine $\underline{\theta}(\cdot, T)$ is thus of the form

$$\rho h_0 \int_0^b \left(\alpha_1\theta_1(r, T) + \alpha_3\theta_3(r, T)\right)\, r dr = \int_0^b \left(J_{\alpha_1}\alpha_1 + J_{\alpha_1'}\alpha_1' + J_{\alpha_3'}\alpha_3' \right)\, r dr. \tag{8.3.11}$$

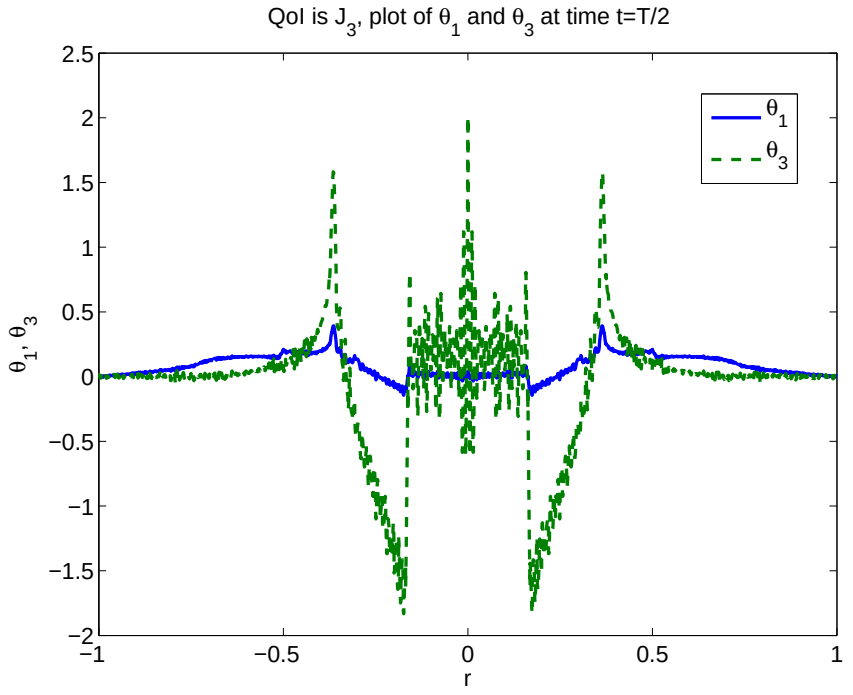
This relation needs to hold for all appropriate α_1 and α_3 and the difficulty is that there are terms in α_1' and α_3' on the right hand side but not on the left hand side. We cannot use integration by parts as $J_{\alpha_1'}$ and $J_{\alpha_3'}$ are in terms of the finite element type function $\underline{u}_h^m$ are they are not smooth enough on $0 \le r \le 1$. We could only consider this case if the function $\underline{u}_h^m$ is at least continuously differentiable.

The dual solution $\underline{z}_h^m$ that we obtain depends on the function $\underline{U}_h^m$ used as data and the functional J being considered. There is no obvious physical interpretation of the quantities

$\psi_1(r,t)$, $\psi_3(r,t)$, $\theta_1(r,t)$ and $\theta_3(r,t)$ but it is interesting to show what is obtained during the computation. In the case of the most accurate approximation obtained in tables 8.3.50 and 8.3.51 we show in figures 8.5(a) and 8.5(b) the profiles at the half-way stage time $t = T/2$ in the case of J_3 and we show in figures 8.6(a) and 8.6(b) the profiles at the half-way stage time $t = T/2$ in the case of J_4. There is highly oscillatory behaviour for θ_1 and θ_3 in both cases but this does not appear much in the expression $F(\underline{z}_h^m) - A(\underline{U}; \underline{z}_h^m)$ as a possible explanation why the estimate is still good in approximating the true error.

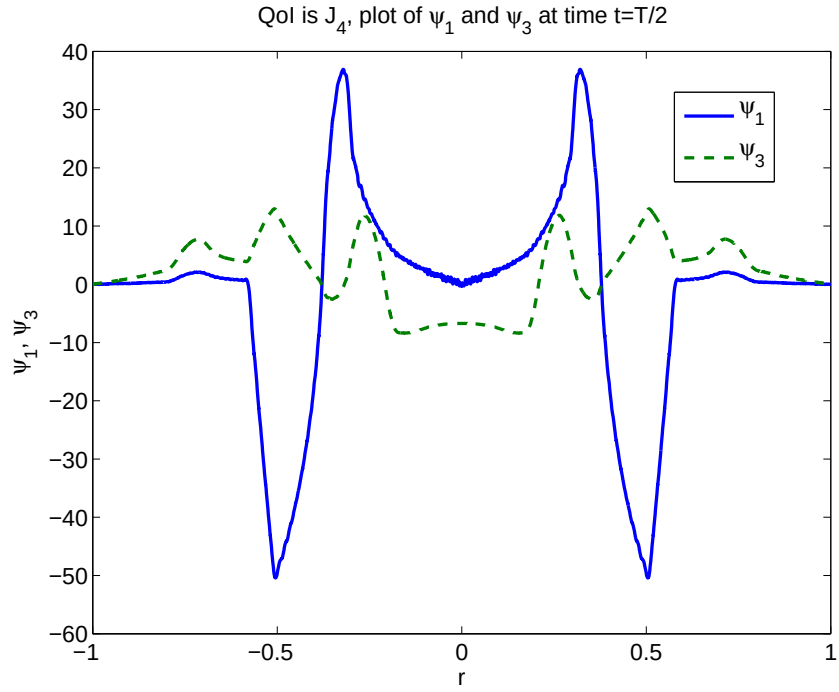


(a) A graph of $\psi_1(r, T/2)$ and $\psi_3(r, T/2)$ with J_3

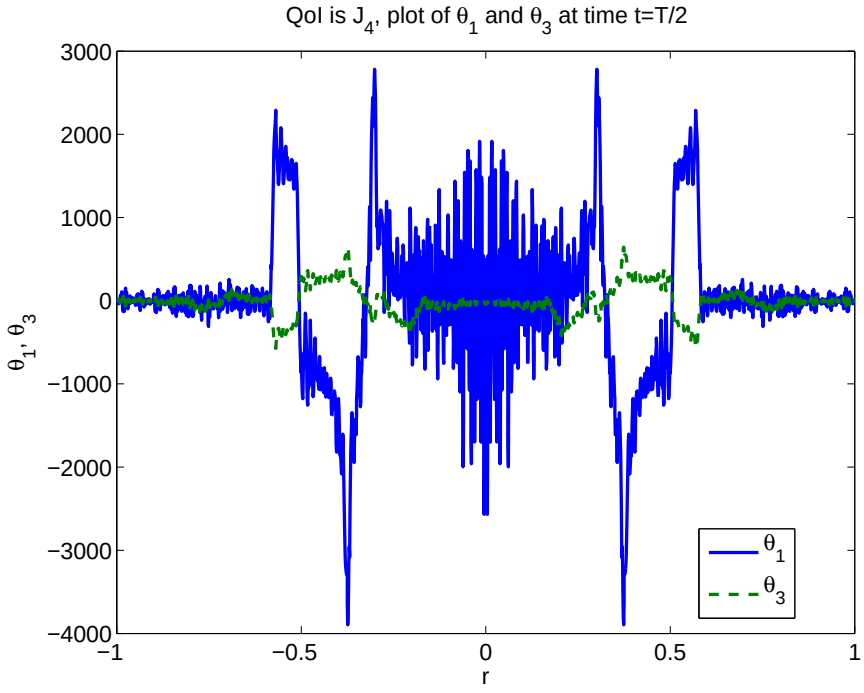


(b) A graph of $\theta_1(r, T/2)$ and $\theta_3(r, T/2)$ with J_3

Fig. 8.5: This is for the functional J_3 with profiles being of $\psi_1(r, T/2)$, $\psi_3(r, T/2)$ in the top figure and it is of $\theta_1(r, T/2)$, $\theta_3(r, T/2)$, $0 \leq r \leq 1$ in the bottom figure.



(a) A graph of $\psi_1(r, T/2)$ and $\psi_3(r, T/2)$ with J_4



(b) A graph of $\theta_1(r, T/2)$ and $\theta_3(r, T/2)$ with J_4

Fig. 8.6: This is for the functional J_4 with profiles being of $\psi_1(r, T/2)$, $\psi_3(r, T/2)$ in the top figure and it is of $\theta_1(r, T/2)$, $\theta_3(r, T/2)$, $0 \leq r \leq 1$ in the bottom figure.

9. CONCLUSIONS

The objectives of this work were to apply the error estimation technique to physical problems involving the inflation of a thin sheet which is assumed to satisfy a membrane model. In each physical problem considered the equations describing the problem are written in a weak form and an approximate solution is obtained using the finite element method. As the physical problems have involved a large deformation, the weak form description, which we write as

$$A(\underline{u};\underline{\psi}) = F(\underline{\psi}), \quad \forall \underline{\psi} \in V, \tag{9.0.1}$$

is nonlinear with $A(\cdot;\cdot)$ denoting a semi-linear form, $F(\cdot)$ denoting a linear functional and with V being an appropriate function space. If for the description here $\underline{u}$ denotes the exact solution and $J(\underline{u})$ denotes the quantity of interest we wish to compute then our estimate is $J(\underline{u}_h)$, where $\underline{u}_h$ is our approximation to $\underline{u}$, and the error estimation technique considered throughout the thesis has involved setting up a related dual problem which uses $\underline{u}_h$ as data. The dual problem to solve is always linear and the dual solution $\underline{z}$ that we obtain gives us an estimate of the error of the form

$$J(\underline{u}) - J(\underline{u}_h) \approx F(\underline{z}) - A(\underline{u}_h;\underline{z}). \tag{9.0.2}$$

As we have described, there are in fact slightly different computational dual problems that can be considered with the easiest one to set-up being of the following form. Find $\underline{z} \in \bar{V}_h$ such that

$$A'(\underline{u}_h;\underline{\alpha},\underline{z}) = J'(\underline{u}_h;\underline{\alpha}) \quad \forall \underline{\alpha} \in \bar{V}_h, \tag{9.0.3}$$

where $\bar{V}_h$ is appropriate finite element space which is different to the space used to get $\underline{u}_h$. When a better approximation $\underline{u}_h^b$ of $\underline{u}$ can be obtained and we let $\underline{u}_h^m = (\underline{u}_h + \underline{u}_h^b)/2$ then the $\underline{z}$ which satisfies

$$A'(\underline{u}_h^m;\underline{\alpha},\underline{z}) = J'(\underline{u}_h^m;\underline{\alpha}) \quad \forall \underline{\alpha} \in \bar{V}_h, \tag{9.0.4}$$

leads to a better estimate in (9.0.2) in general. In the thesis we have demonstrated that this abstract framework can be applied to the membrane inflation problems described, provided we have a sufficiently close approximation $\underline{u}_h$ to $\underline{u}$. Specifically, when $\underline{z}$ satisfies (9.0.4) the theory only indicates that the estimate in (9.0.2) will be good when $\underline{u}_h$ is sufficiently close to $\underline{u}$ and for any given problem which is being considered for the first time it is not clear in advance how close $\underline{u}_h$ needs to be to $\underline{u}$ and this has been an issue in the problems considered. When the physical problems involve a quasi-static deformation and the unknown displacement just depends on space variables, everything works quite well even with relatively coarse meshes. When the physical problem involves the equations of motion the scheme also works but only after sufficient effort has been put into the time discretization. In fact, in the examples, we frequently do not get any reliable estimate of the error in the dynamic case until the approximate solution is sufficiently accurate. The effort needed for how the approximate solution varies in time is one of the more surprising conclusions from the study although when we are attempting to estimate the error in a computation it is always going to be the case that this is likely to be dominated by the least accurate part of the overall procedure. In summary we have been successful in demonstrating that the error estimation technique does work for each of the three problems considered although for future work ways to reduce the amount of computation to estimate $J(\underline{u}) - J(\underline{u}_h)$ should be investigated. At present the larger space $\bar{V}_h$ used to get a dual solution has involved piecewise polynomials of one degree higher that used to get $\underline{u}_h$ and hence solving the dual problem typically involves more computational effort than is used to get $\underline{u}_h$ and $J(\underline{u}_h)$. Perhaps higher degree polynomials but on a coarser mesh might be a possibility to try for $\bar{V}_h$ although this would complicate a little the implementation with different meshes being involved at the same stage of the procedure.

Two of the three problems considered just involved a quasi-static deformation and a displacement field $\underline{u} = \underline{u}(x_1, x_2)$ or $\underline{u} = \underline{u}(r)$ which only depends on space variables at a given applied pressure $P = P(t)$. There is detail to cope with to get the Gâteaux derivative $A'(\cdot;\cdot,\cdot)$ of the semi-linear form $A(\cdot;\cdot)$ in the weak form and care is needed in the expressions when the two principal stretches λ_1 and λ_2 are the same but otherwise there are not too many difficulties in applying the technique. This is partly because we have kept to pressures low enough that we do not have a limit point in the nonlinear system at which the Jacobian matrix at the solution is singular. The entries in the Jacobian matrix can actually be expressed in terms of $A'(\cdot;\cdot,\cdot)$, i.e. the expression for $A'(\cdot;\cdot,\cdot)$ appears in linear dual problems and in the Jacobian matrix and hence we effectively only need to get the expressions correct once in an implementation. In the examples considered the displacement $\underline{u}$ as result of the geometry of Ω and the pressure loading is mostly

quite smooth and when using high degree polynomials is straightforward we can quite easily get very accurate answers. There are several examples in the axisymmetric case to demonstrate this and adaptive refinement based on element quantities of the form

$$F(\underline{z} - \underline{z}_I)_k - A(\underline{u}_h; \underline{z} - \underline{z}_I)_k, \quad k = 1, \ldots, ne \tag{9.0.5}$$

works very well. For example, if we attempt to estimate by how much we need to refine to reach a given accuracy then this is usually successful in getting the accuracy in one or two steps. The axisymmetric quasi-static inflation problem is however only one dimensional and it is the easiest of the problems considered. In the non-axisymmetric case things do not work quite as well although with piecewise linears on triangular meshes to get $\underline{u}_h$ and with quadratics on the same triangles to get the dual solution does lead to an accurate estimate of the error. The less clear aspect in the non-axisymmetric case is in deciding which elements to refine based on the element indicators in (9.0.5) which are typically not all of the same sign. In the examples in chapter 5 the quantities of interest considered, have usually suggested that uniform refinement should be done and we only have one example involving a L-shape where we have one non-uniform refinement step.

The dynamic problem is theoretically the most difficult of the three problems in the thesis which in the axisymmetric case involves a space time region

$$\{(r,t) : 0 \le r \le 1,\, 0 \le t \le T\}\,.$$

In the numerical scheme we have time levels $0 = t_0 < t_1 < \cdots t_M = T$ and we solve forward to time to get the approximate displacement and velocity on each time interval $[t_{j-1}, t_j]$, $j = 1, \ldots, M$. The related dual problem involves solving backwards in time, i.e. we start with getting the solution at time $t = t_M = T$ and then we obtain $\underline{z}(r,t)$, $t \in [t_{j-1}, t_j]$, $j = M, \ldots, 2, 1$. To improve the accuracy in a quantity of interest by a sufficient amount we need, at some stage, to refine in both space and time in some way although we need to first determine that stage. Before that stage is reached we may be able to reduce the error by just refining in space or with just refining in time depending on the problem being considered. Also, when we are at the stage that the error is decreasing, the rate of the decrease will be determined by the least accurate part of the approximation. It is this last observation that led us to try higher order schemes for the time dependence of the approximation on each time interval $[t_{j-1}, t_j]$. This is moderately complicated in the detail but it becomes more manageable once it is noted that when degree n polynomials in time are used on $[t_{j-1}, t_j]$ the difference between the approximate velocity $\underline{v}_h(r,t)$ and $\dot{\underline{u}}_h(r,t)$ involves the Legendre polynomial of degree n on $[t_{j-1}, t_j]$ with the scheme used. With a way of using any degree p in space and any degree n in time there is a lot of choice

as to what to try get both an accurate approximation and a reliable estimate of the error in the approximation. In the thesis we do not reach the stage of determining the optimum choice for p and n in terms of reaching a desired accuracy, with a reliable estimate of the error with the least computational effort. However, we seem to be able to do quite well in the examples with the space degree $p = 2$ and with the time degree of $n \geq 3$.

In the larger context of modelling how thin sheets deform in an industrial process known as thermoforming it should be appreciated which parts of the work described here can be used and which parts cannot. To be able to use the technique described in this thesis we need to be able to write the problem in a weak form and as presented this excludes contact problems which is a key aspect of a forming process. To be able to express the problem in a weak form may also restrict the constitutive model that can be used if constitutive models other that hyperelastic models are to be considered. There is no obvious reason however why we cannot consider the body as a general three dimensional solid or to have a non-homogeneous body. The accuracy to attempt to achieve is such cases is likely to be influenced by the accuracy for which the material data is known.

BIBLIOGRAPHY

[1] M. Ainsworth and J.T. Oden. *A Posteriori Error Estimation in Finite Element Analysis*. A Wiley-Interscience publication. Wiley, 2000.

[2] GD Akribis. *Finite Element Methods*. University Publication, University of Cyprus, 2005.

[3] H. Andrà, M. Warby, and J.R. Whiteman. Contact problems of hyperelastic membranes: existence theory. *Mathematical Methods in the Applied Sciences*, 23(10):865–895, July 2000.

[4] R.J. Atkin and N. Fox. *An Introduction to the Theory of Elasticity*. Dover Books on Physics. Dover Publications, 2013.

[5] I. Babuska, J. Whiteman, and T. Strouboulis. *Finite Elements: An Introduction to the Method and Error Estimation*. OUP Oxford, 2010.

[6] Wolfgang Bangerth and Rolf Rannacher. Finite element approximation of the acoustic wave equation: Error control and mesh adaptation. *East West Journal of Numerical Mathematics*, 7(4):263–282, 1999.

[7] R.E. Bank and A Weiser. Some a posteriori error estimators for elliptic partial differential equations. *Math. Comp.*, 44:283–301, 1985.

[8] Roland Becker and Rolf Rannacher. An optimal control approach to a posteriori error estimation in finite element methods. *Acta Numerica 2001*, 10:1–102, 2001.

[9] Ted Belytschko, Wing Kam Liu, Brian Moran, and Khalil Elkhodary. *Nonlinear finite elements for continua and structures*. John Wiley & Sons, 2013.

[10] S. Brenner and R. Scott. *The Mathematical Theory of Finite Element Methods*. Texts in Applied Mathematics. Springer New York, 2000.

[11] François Clément and Vincent Martin. The lax-milgram theorem. A detailed proof to be formalized in coq. *CoRR*, abs/1607.03618, 2016.

[12] A.D. Drozdov. *Finite Elasticity and Viscoelasticity: A Course in the Nonlinear Mechanics of Solids*. World Scientific, 1996.

[13] A.E. Green and J.E. Adkins. *Large elastic deformations*. Clarendon Press, 1970.

[14] P. Haupt. *Continuum mechanics and theory of materials*. Springer, 2002.

[15] C. Johnson. *Numerical Solution of Partial Differential Equations by the Finite Element Method*. Cambridge U. Press, 1987.

[16] David Kincaid and Ward Cheney. *Numerical Analysis: Mathematics of Scientific Computing*. Brooks/Cole Publishing Co., Pacific Grove, CA, USA, 1991.

[17] J Tinsley Oden and Serge Prudhomme. Estimation of modeling error in computational mechanics. *Journal of Computational Physics*, 182(2):496–515, 2002.

[18] J Tinsley Oden, Serge Prudhomme, Daniel C Hammerand, and Mieczyslaw S Kuczma. Modeling error and adaptivity in nonlinear continuum mechanics. *Computer Methods in Applied Mechanics and Engineering*, 190(49):6663–6684, 2001.

[19] J Tinsley Oden and Kumar S Vemaganti. Estimation of local modeling error and goal-oriented adaptive modeling of heterogeneous materials: I. error estimates and adaptive algorithms. *Journal of Computational Physics*, 164(1):22–47, 2000.

[20] John Tinsley Oden and Serge Prudhomme. Goal-oriented error estimation and adaptivity for the finite element method. *Computers & Mathematics with Applications*, 41(5):735–756, 2001.

[21] R. W. Ogden. Large deformation isotropic elasticity - on the correlation of theory and experiment for incompressible rubberlike solids. *Proceedings of the Royal Society A*, 326(1567):565–584, 1972.

[22] Serge Prudhomme and J Tinsley Oden. On goal-oriented error estimation for elliptic problems: application to the control of pointwise errors. *Computer Methods in Applied Mechanics and Engineering*, 176(1):313–331, 1999.

[23] Serge Prudhomme, J Tinsley Oden, Tim Westermann, Jon Bass, and Mark E Botkin. Practical methods for a posteriori error estimation in engineering applications. *International Journal for Numerical Methods in Engineering*, 56(8):1193–1224, 2003.

[24] R. S. Rivlin. Large Elastic Deformations of Isotropic Materials. IV. Further Developments of the General Theory. *Philosophical Transactions of the Royal Society of London Series A*, 241:379–397, October 1948.

[25] R. S. L Rivlin. Large elastic deformations of isotropic materials. i. fundamental concepts. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 240(822):459–490, 1948.

[26] Ronald S Rivlin and DW Saunders. Large elastic deformations of isotropic materials. vii. experiments on the deformation of rubber. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 243(865):251–288, 1951.

[27] Walter Rudin. *Real and Complex Analysis, 3rd Ed.* McGraw-Hill, Inc., New York, NY, USA, 1987.

[28] S Shaw, MK Warby, and JR Whiteman. Discretization error and modelling error in the context of the rapid inflation of hyperelastic membranes. *IMA Journal of Numerical Analysis*, 30(1):302–333, 2010.

[29] A.J.M. Spencer. *Continuum Mechanics*. Dover Books on Physics. Dover Publications, 2012.

[30] Barna Szabó and Ivo Babuška. *Introduction to finite element analysis: formulation, verification and validation*, volume 35. John Wiley & Sons, 2011.