MODELLING AND NUMERICAL SIMULATION IN CONTINUUM MECHANICS

Coimbra, July 14-18, 2003

Isabel Narra Figueiredo, Luís Filipe Menezes, Juha Videman

23

CENTRO INTERNACIONAL de MATEMÁTICA

The Advanced School and Workshop on Modelling and Numerical Simulation in Continuum Mechanics was one of the events included in the C.I.M 2003 Thematic Term in Mathematics and Engineering. The event consisted of an interdisciplinary school and of an application-orientated workshop. It was held at the Department of Mechanical Engineering of the Faculty of Sciences and Technology of the University of Coimbra, from July 14 to 18, 2003.

Twenty two speakers gave lectures at the event presenting either short courses, plenary lectures, or invited/contributed talks. Various problems in Structural Mechanics, Shell Theory, Shape Optimization and Fluid Mechanics were addressed either from modelling, computational or industrial application viewpoints. In this Volume, we have gathered both the lecture notes, written for the short courses and the plenary lectures, and the one page abstracts of the invited and contributed talks.

On behalf of the C.I.M (Centro International de Matemática) we express our most sincere thanks to all the speakers for their valuable communications. We are also grateful for the generous financial support from Fundação Calouste Gulbenkian, CMUC (Centro de Matemática da Universidade de Coimbra), CEMUC (Centro de Engenharia Mecânica da Universidade de Coimbra), CEMAT-IST (Centro de Matemática e Aplicações do Instituto Superior Técnico) and FCT (Fundação para a Ciência e a Tecnologia).

Coimbra, July 2003

Organizing Committee

Luís Filipe Menezes (Universidade de Coimbra)
Isabel Narra de Figueiredo (Universidade de Coimbra)
Juha Hans Videman (Instituto Superior Técnico)

CONTENTS

Contributed Talks

- Ángel Rodríguez-Arós, M. Sofonea, J.M. Viaño : *Some contact problems in viscoelasticity with long-term memory*
- Cristian Barbarosie : *Analysis and optimization of mixtures of materials at the microscale level*
- M. Campo, J.R. Fernández, M. Shillor : *A viscoelastic beam oscillating between two stops with damage*
- A. Pinto da Costa, J.A.C. Martins : *Instability and bifurcation modes in systems with Coulomb friction*
- Carolina Ribeiro, Juan M. Viaño : *Link between the Kirchhoff-Love and the Bernoulli-Navier models for a rectangular plate/beam*
- Lourenço Beirão da Veiga : *A new numerical scheme for Von-Mises plasticity with linear hardening*

Numerical analysis of discrete schemes approximating grade-two fluid models. Recent results and open problems

V. Girault*

March 31, 2003

Abstract

These notes are devoted to some numerical schemes for approximating the solutions of incompressible grade-two fluid models in 2-D. First, we recall briefly the essential points of the theoretical analysis of the model. Next we take advantage of the information gained through this analysis to devise appropriate numerical schemes and algorithms. We include considerations on a scheme for the 3-D model, whose numerical analysis is still an open problem.

1 Introduction

A grade-two fluid is one of the theoretical models introduced by Rivlin and Ericksen [55] for describing non-Newtonian behaviour. Its equations generalize the Navier-Stokes equations and it is believed to describe the motion of a water solution of polymers (cf. [23]). Interestingly, its equations have been interpreted recently by Camassa, Holm, Marsden, Ratiu and Shkoller (cf. for instance [36, 37]) as a model of turbulence. In the simplest case, the equations have the form

$$\frac{\partial}{\partial t}(\boldsymbol{u} - \alpha\Delta\,\boldsymbol{u}) - \nu\Delta\,\boldsymbol{u} + \mathbf{curl}(\boldsymbol{u} - \alpha\Delta\,\boldsymbol{u}) \times \boldsymbol{u} + \nabla p = \boldsymbol{f}\,, \tag{1.1}$$

$$\operatorname{div}\boldsymbol{u} = 0\,, \tag{1.2}$$

with tangential Dirichlet boundary conditions and an initial condition. Here $\boldsymbol{u}$ is the velocity, p is related to the pressure, $\nu \geq 0$ is the viscosity and $\alpha > 0$ is the normal stress-modulus when the equations model a non-Newtonian fluid, and an averaged-length scale when the equations model turbulence.

Analyzing schemes approximating a grade-two fluid model is interesting, not only on account of these two interpretations, but also because it gives an insight on what can be done to overcome the difficulties raised by the highly non-linear term $\mathbf{curl}\,\Delta\,\boldsymbol{u} \times \boldsymbol{u}$, and what are the open problems if they have not yet been overcome.

In some sense, the theoretical results that have been proven up to date for this model are fairly satisfactory, but there still remain important open questions such as the problem posed by non-homogeneous Dirichlet boundary conditions or that posed by a rough exterior force, such as an L^2 force, to mention just these two "simple" questions. At least for the steady 2-D problem, we can handle tangential Dirichlet boundary conditions, i.e. with no ingoing or outgoing flow. But if there is an ingoing or outgoing flow, the problem is ill-posed and we do not know what additional boundary condition must be added to make the problem well-posed.

In contrast, results of numerical analysis obtained so far are very scanty. We now know how to analyze carefully chosen schemes for the steady problem in 2-D. Again in 2-D, we can hopefully do the numerical

*Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie, Paris, France, email: girault@ann.jussieu.fr

analysis of schemes that approximate the time-dependent problem without expecting major difficulties. But so far, no one knows how to analyze schemes that approximate this problem in 3-D, be it steady or unsteady. The explanation is simple: we lack some discrete a priori estimates, estimates that appear plausible, but for which we have yet no proof. These estimates are a crucial ingredient in the numerical analysis of several models of non-Newtonian fluids, and this analysis will remain an open question as long as these estimates are not established.

In 2-D, a grade-two fluid model has a global solution without restriction on the size of the data and on the boundary of the domain, exactly as for the Navier-Stokes equations. This remarkable property is due to the fact that, when the problem is written in the form of a generalized Stokes equation and a transport equation, solutions can be constructed without requiring that the velocity be bounded in $W^{1,\infty}$. Then if the equations of a grade-two fluid model are suitably discretized, and the finite-element spaces well-chosen, this property can be preserved. In this case, the numerical analysis of such schemes can be done successfully. This analysis does not carry over yet to 3-D because the exact problem requires a velocity in $W^{1,\infty}$ and hence the discrete schemes also require that the discrete velocity be uniformly bounded in $W^{1,\infty}$, with respect to the discretization parameters.

After this introduction, we recall the analysis of the exact problem (1.1), (1.2). Then we present a centered and an upwind scheme in 2-D, for which we can establish a priori estimates, existence of discrete solutions and their strong convergence without restriction on the data and the domain. Then, by suitably restricting the domain and the data, we prove uniqueness and error estimates and establish the convergence of a simple algorithm for computing the discrete solution. We propose a scheme for the 3-D model and examine the open questions raised by its numerical analysis. Finally, we describe a least-squares algorithm that gives good results, but for which there is no analysis.

We close this introduction with a list of notation that will be used in the sequel. We state them in 3-D because the theoretical problem is of course three-dimensional, but the numerical study will be done mainly in 2-D. Unless otherwise specified, the domains of interest Ω will all have a boundary $\partial\Omega$ that is at least Lipschitz-continuous (cf. [35]). Let (k_1, k_2, k_3) be a triple of non-negative integers and set $|k| = k_1 + k_2 + k_3$; we define the partial derivative ∂^k of order k:

$$\partial^k v = \frac{\partial^{|k|} v}{\partial x_1^{k_1} \partial x_2^{k_2} \partial x_3^{k_3}} \cdot$$

Recall the standard Sobolev spaces, for a non-negative integer m and a number $r \geq 1$ (cf. [1] or [46])

$$W^{m,r}(\Omega) = \{ v \in L^r(\Omega) \,;\, \partial^k v \in L^r(\Omega) \ \forall |k| \leq m \} \,,$$

equipped with the seminorm

$$|v|_{W^{m,r}(\Omega)} = \left[\sum_{|k|=m} \int_\Omega |\partial^k v|^r \, d\boldsymbol{x} \right]^{1/r} ,$$

and the norm (for which it is a Banach space)

$$\|v\|_{W^{m,r}(\Omega)} = \left[\sum_{0 \leq k \leq m} |v|^r_{W^{k,r}(\Omega)} \right]^{1/r} ,$$

with the usual modification when $r = \infty$; we refer to [35], [45] or [1] for extending this definition to fractional Sobolev spaces. When $r = 2$, this space is the Hilbert space $H^m(\Omega)$. In particular, the scalar product of $L^2(\Omega)$ is denoted by $(\cdot, \cdot)$. These definitions are extended straightforwardly to vector-valued functions, with the same notation, except for non-Hilbert norms. In the case of a vector $\boldsymbol{u} = (u_1, u_2, u_3)$, we set

$$\|\boldsymbol{u}\|_{L^r(\Omega)} = \left[\int_\Omega |\boldsymbol{u}(\boldsymbol{x})|^r \, d\boldsymbol{x} \right]^{1/r} ,$$

where $|\cdot|$ denotes the Euclidian norm: $|\boldsymbol{u}|^2 = \boldsymbol{u}\cdot\boldsymbol{u}$.

For imposing vanishing boundary values on $\partial\Omega$, we define

$$H_0^1(\Omega) = \{v \in H^1(\Omega)\,;\, v|_{\partial\Omega} = 0\}\,.$$

We shall frequently use Sobolev imbeddings: for a real number $p \geq 1$ in 2-D or $1 \leq p \leq 6$ in 3-D, there exists a constant S_p (that depends only on the dimension and the domain) such that

$$\forall v \in H_0^1(\Omega)\ ,\ \|v\|_{L^p(\Omega)} \leq S_p |v|_{H^1(\Omega)}\,. \tag{1.3}$$

When $p = 2$, this is Poincaré's inequality and S_2 is Poincaré's constant. Owing to Poincaré's inequality, the seminorm $|\cdot|$ is a norm on $H_0^1(\Omega)$, equivalent to the full norm. As it it is directly related to the gradient operator, we choose this seminorm as norm on $H_0^1(\Omega)$, and in particular, we use it to define the dual norm on its dual space $H^{-1}(\Omega)$.

For imposing tangential boundary conditions, we define

$$H_T^1(\Omega) = \{\boldsymbol{v} \in H^1(\Omega)^3\,;\, \boldsymbol{v}\cdot\boldsymbol{n} = 0 \text{ on } \partial\Omega\}\,, \tag{1.4}$$

where $\boldsymbol{n}$ est the unit normal vector to $\partial\Omega$, directed outside Ω. An easy application of Peetre-Tartar's Theorem (cf. [50], [59] or [29]) proves the analogue of Sobolev's imbeddings in $H_T^1(\Omega)$ for any real number $p \geq 1$ in 2-D or $1 \leq p \leq 6$ in 3-D:

$$\forall \boldsymbol{v} \in H_T^1(\Omega)\ ,\ \|\boldsymbol{v}\|_{L^p(\Omega)} \leq \tilde{S}_p |\boldsymbol{v}|_{H^1(\Omega)}\,. \tag{1.5}$$

In particular, for $p = 2$, the mapping $\boldsymbol{v} \mapsto |\boldsymbol{v}|_{H^1(\Omega)}$ is a norm on $H_T^1(\Omega)$, equivalent to the H^1 norm and $\tilde{S}_2$ is the analogue of Poincaré's constant. We shall also use the classical spaces for Navier-Stokes equations:

$$V = \{\boldsymbol{v} \in H_0^1(\Omega)^3\,;\, \operatorname{div}\boldsymbol{v} = 0 \text{ in } \Omega\}\ ,\ \operatorname{div}\boldsymbol{v} = \sum_{i=1}^{3} \frac{\partial v_i}{\partial x_i}\,, \tag{1.6}$$

$$W = \{\boldsymbol{v} \in H_T^1(\Omega)\,;\, \operatorname{div}\boldsymbol{v} = 0 \text{ in } \Omega\}\,, \tag{1.7}$$

$$L_0^2(\Omega) = \{q \in L^2(\Omega)\,;\, \int_\Omega q\, d\boldsymbol{x} = 0\}\,,$$

$$H(\mathbf{curl}, \Omega) = \{\boldsymbol{v} \in L^2(\Omega)^3\,;\, \mathbf{curl}\,\boldsymbol{v} \in L^2(\Omega)^3\}\,,$$

where

$$\mathbf{curl}\,\boldsymbol{v} = \big(\frac{\partial v_3}{\partial x_2} - \frac{\partial v_2}{\partial x_3}, \frac{\partial v_1}{\partial x_3} - \frac{\partial v_3}{\partial x_1}, \frac{\partial v_2}{\partial x_1} - \frac{\partial v_1}{\partial x_2}\big)\,. \tag{1.8}$$

2 Formulation

A grade-two fluid is a fluid of differential type. Its Cauchy stress $\boldsymbol{T}$ is given by:

$$\boldsymbol{T} = -p\boldsymbol{I} + \mu\boldsymbol{A}_1 + \alpha_1\boldsymbol{A}_2 + \alpha_2\boldsymbol{A}_1^2\,, \tag{2.1}$$

where $\boldsymbol{I}$ is the identity tensor and $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ are the first two Rivlin-Ericksen tensors [55] defined recursively by

$$\boldsymbol{A}_1 = \boldsymbol{L}_1 + \boldsymbol{L}_1^t\ ,\ (\boldsymbol{L}_1)_{i,j} = (\nabla\,\boldsymbol{u})_{i,j} = \frac{\partial u_i}{\partial x_j}\ ,\ \boldsymbol{A}_2 = \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{A}_1 + \boldsymbol{A}_1\boldsymbol{L}_1 + \boldsymbol{L}_1^t\boldsymbol{A}_1\,.$$

Here $\frac{\mathrm{d}}{\mathrm{d}t}$ denotes the convective or material derivative :

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{A}_1 = \frac{\partial}{\partial t}\boldsymbol{A}_1 + \boldsymbol{u}\cdot\nabla\boldsymbol{A}_1\,.$$

For the fluid to be thermodynamically compatible (cf. [22]), the viscosity μ and the material moduli α_1 and α_2 must satisfy

$$\mu \geq 0 \ , \ \alpha_1 \geq 0 \ , \ \alpha_1 + \alpha_2 = 0 \, . \tag{2.2}$$

By substituting (2.1) into the balance of linear momentum:

$$\varrho \frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{u} = \operatorname{div} \boldsymbol{T} + \varrho \boldsymbol{f} \, , \tag{2.3}$$

where ϱ is the density and $\boldsymbol{f}$ is an exterior force (such as gravity) and dividing by the density, we find the momentum equation of a grade-two fluid:

$$\begin{aligned} \frac{\partial}{\partial t}(\boldsymbol{u} - \alpha_1 \Delta \boldsymbol{u}) - \mu \, \Delta \boldsymbol{u} + \mathbf{curl}(\boldsymbol{u} - (2\alpha_1 + \alpha_2)\Delta \boldsymbol{u}) \times \boldsymbol{u} \\ - (\alpha_1 + \alpha_2)\Delta(\boldsymbol{u} \cdot \nabla \boldsymbol{u}) + 2(\alpha_1 + \alpha_2)\boldsymbol{u} \cdot \nabla(\Delta \boldsymbol{u}) + \nabla p = \boldsymbol{f} \, . \end{aligned} \tag{2.4}$$

To simplify, we keep the same notation for the parameters and for the pressure:

$$\mu := \frac{\mu}{\varrho} \, , \, \alpha_i := \frac{\alpha_i}{\varrho} \ , \ p := \frac{1}{\varrho} p + \frac{1}{2}|\boldsymbol{u}|^2 - (2\,\alpha_1 + \alpha_2)(\boldsymbol{u} \cdot \Delta \, \boldsymbol{u} + \frac{1}{4} \operatorname{tr} \boldsymbol{A}_1^2) \, .$$

As $\alpha_1 + \alpha_2 = 0$, we set $\alpha = \alpha_1$ and (2.4) simplifies to (1.1):

$$\frac{\partial}{\partial t}(\boldsymbol{u} - \alpha \Delta \, \boldsymbol{u}) - \mu \, \Delta \, \boldsymbol{u} + \mathbf{curl}(\boldsymbol{u} - \alpha \Delta \, \boldsymbol{u}) \times \boldsymbol{u} + \nabla \, p = \boldsymbol{f} \, ,$$

that must be completed by the incompressibility condition (1.2), an initial condition at time $t = 0$ and a no-slip boundary condition.

Remark 2.1 Note that when $\alpha = 0$, (1.1) reduces to the Navier-Stokes equations owing to the identity:

$$\boldsymbol{u} \cdot \nabla \, \boldsymbol{u} = \mathbf{curl} \, \boldsymbol{u} \times \boldsymbol{u} + \frac{1}{2} \nabla (|\boldsymbol{u}|^2) \, .$$

As far as the steady 2-D problem is concerned, we prove further on that when α tends to zero, the corresponding solutions tend to solutions of the Navier-Stokes problem. This is possibly also true in 3-D, but it seems unlikely for the evolution problem (see Remark 3.5).

The condition $\alpha \geq 0$ has been (and is still) a source of rough controversy. From a mathematical point of view, the term $-\frac{\partial}{\partial t} \alpha \Delta \, \boldsymbol{u}$ in the left-hand side of the momentum equation makes the rest-state unstable when α is negative, and therefore, we shall not study this case here. ∎

3 Theoretical analysis

Let Ω be a bounded domain of $I\!R^3$, with a Lipschitz boundary $\partial\Omega$. Consider the problem: Find a velocity vector $\boldsymbol{u}$ and a scalar pressure p, solution of

$$\frac{\partial}{\partial t}(\boldsymbol{u} - \alpha \Delta \, \boldsymbol{u}) - \mu \Delta \, \boldsymbol{u} + \mathbf{curl}(\boldsymbol{u} - \alpha \Delta \, \boldsymbol{u}) \times \boldsymbol{u} + \nabla \, p = \boldsymbol{f} \quad \text{in } \Omega \times]0, T[\, , \tag{3.1}$$

with the incompressibility condition:

$$\operatorname{div} \boldsymbol{u} = 0 \quad \text{in } \Omega \times]0, T[\, , \tag{3.2}$$

to simplify, we only impose a homogeneous Dirichlet boundary condition:

$$\boldsymbol{u} = \boldsymbol{0} \quad \text{on } \partial\Omega \times]0, T[\, , \tag{3.3}$$

and the initial condition:

$$\boldsymbol{u}(0) = \boldsymbol{u}_0 \text{ in } \Omega \ \text{ with } \operatorname{div} \boldsymbol{u}_0 = 0 \text{ in } \Omega \text{ and } \boldsymbol{u}_0 = \boldsymbol{0} \text{ on } \partial\Omega \, . \tag{3.4}$$

Remark 3.1 Considering that (3.1) involves a third derivative, we can ask the question: does (3.3) impose enough boundary conditions to determine the solution of (3.1)–(3.4)? We shall see further on that the answer is "yes". More generally, [30] proves that the answer is also "yes" for the steady-state problem in 2-D in the case where (3.3) is replaced by a tangential Dirichlet condition:

$$\boldsymbol{u} = \boldsymbol{g} \quad \text{on } \partial\Omega\times]0,T[\quad \text{with} \quad \boldsymbol{g}\cdot\boldsymbol{n} = 0\,. \tag{3.5}$$

It is likely that, with adequate conditions on $\boldsymbol{g}$, this result extends to the evolution problem (3.1)–(3.4). But when the boundary values are not tangential, there are examples where the problem is ill-posed, cf. [54]. ∎

Problem (3.1)–(3.4) is difficult because its non-linear term involves a third order derivative, whereas its elliptic part only comes from a Laplace operator; for this reason, it behaves mostly as a hyperbolic problem. For the past ten years, many publications have been devoted to this problem, but by far the best proof of existence, due to Cioranescu and Ouazar, goes back to more than twenty years ago (1981) and is found in the thesis of Ouazar [48]; it was published later by Cioranescu and Ouazar in [17, 18].

Here is a brief description of their construction. Some of its ideas will be very valuable for discretizing the problem. Their assumptions on the data and the domain are: Ω simply-connected of class $\mathcal{C}^{3,1}$, $\boldsymbol{f}$ in $L^2(0,T;H^1(\Omega)^3)$ and $\boldsymbol{u}_0$ in $H^3(\Omega)^3$. Formally, observe first that (3.1) yields the energy equality:

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{u}(t)\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\frac{\mathrm{d}}{\mathrm{d}t}|\boldsymbol{u}(t)|^2_{H^1(\Omega)} + \mu|\boldsymbol{u}(t)|^2_{H^1(\Omega)} = (\boldsymbol{f}(t),\boldsymbol{u}(t))\,. \tag{3.6}$$

It shows in particular that, if a solution $\boldsymbol{u}$ exists, then it is unconditionally bounded in $L^\infty(0,T;H^1(\Omega)^3)$ by the data $\boldsymbol{f}$. Now, set

$$\boldsymbol{z} = \mathbf{curl}(\boldsymbol{u} - \alpha\,\Delta\,\boldsymbol{u})\,. \tag{3.7}$$

This choice is crucial, because Cioranescu and Ouazar prove that if $\mathbf{curl}(\boldsymbol{u} - \alpha\,\Delta\,\boldsymbol{u}) \in L^2(\Omega)^3$ and Ω is simply-connected, then $\boldsymbol{u} \in H^3(\Omega)^3$ and there exists a constant C such that

$$\|\boldsymbol{u}\|_{H^3(\Omega)} \le C\big(|\boldsymbol{u}|^2_{H^1(\Omega)} + \|\mathbf{curl}(\boldsymbol{u} - \alpha\,\Delta\,\boldsymbol{u})\|^2_{L^2(\Omega)}\big)^{1/2}\,. \tag{3.8}$$

Next, take formally the curl of (3.1); this gives a transport equation, (that we multiply here by α):

$$\alpha\frac{\partial}{\partial t}\boldsymbol{z} + \mu\boldsymbol{z} + \alpha\,\boldsymbol{u}\cdot\nabla\,\boldsymbol{z} - \alpha\,\boldsymbol{z}\cdot\nabla\,\boldsymbol{u} = \mu\,\mathbf{curl}\,\boldsymbol{u} + \alpha\,\mathbf{curl}\,\boldsymbol{f} \quad \text{in } \Omega\times]0,T[\,, \tag{3.9}$$

and formally multiply (3.9) by $\boldsymbol{z}$. We obtain the inequality:

$$\frac{\alpha}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{z}(t)\|^2_{L^2(\Omega)} + (\mu - \alpha\|\nabla\,\boldsymbol{u}(t)\|_{L^\infty(\Omega)})\|\boldsymbol{z}(t)\|^2_{L^2(\Omega)} \le (\mu\,\|\mathbf{curl}\,\boldsymbol{u}(t)\|_{L^2(\Omega)} + \alpha\,\|\mathbf{curl}\,\boldsymbol{f}(t)\|_{L^2(\Omega)})\|\boldsymbol{z}(t)\|_{L^2(\Omega)}\,. \tag{3.10}$$

By applying a Sobolev bound to $\|\nabla\,\boldsymbol{u}(t)\|_{L^\infty(\Omega)}$, substituting (3.8) into the left-hand side of (3.10) and the estimate deduced from (3.6) to bound $\|\mathbf{curl}\,\boldsymbol{u}(t)\|_{L^2(\Omega)}$ in its right-hand side, we find that $\|\boldsymbol{z}(t)\|^2_{L^2(\Omega)}$ is bounded by the solution of a Riccati differential equation on the time interval $[0,T^*]$, for some $T^* > 0$, $T^* \le T$. This shows that, if a solution $\boldsymbol{u}$ exists, then it is bounded in $L^\infty(0,T^*;H^3(\Omega)^3)$. Finally, on multiplying formally (3.1) by $\boldsymbol{u}'$ and using the previous bound for $\boldsymbol{u}$, we infer that $\boldsymbol{u}'$ is also bounded in $L^2(0,T^*;H^1(\Omega)^3)$. In principle, these a priori estimates are sufficient to construct a local solution of the problem.

Remark 3.2 On one hand, the energy equality (3.6) explains why it is important that α be positive.

On the other hand, to obtain (3.10), we must eliminate the term $\alpha(\boldsymbol{u}\cdot\nabla\,\boldsymbol{z},\boldsymbol{z})$. In view of (3.2), assuming that Green's formula is valid, we have:

$$\alpha\int_\Omega (\boldsymbol{u}\cdot\nabla\,\boldsymbol{z})\boldsymbol{z}\,d\boldsymbol{x} = \frac{\alpha}{2}\int_{\partial\Omega}(\boldsymbol{u}\cdot\boldsymbol{n})|\boldsymbol{z}|^2 ds\,. \tag{3.11}$$

This term vanishes either if $\boldsymbol{u}\cdot\boldsymbol{n} = 0$ or if $\boldsymbol{z} = \boldsymbol{0}$ where $\boldsymbol{u}\cdot\boldsymbol{n} \ne 0$. In the second case, what is the physical meaning of this condition on $\boldsymbol{z}$? And what is the mathematical meaning of this condition on $\boldsymbol{z}$, when $\boldsymbol{z}$ is only in $L^2(\Omega)^3$, as it is here ? ∎

Remark 3.3 At first sight, (3.6) seems minor compared to (3.8). But in fact, (3.6) is crucial in estimating the term $\|\mathbf{curl}\,\boldsymbol{u}(t)\|_{L^2(\Omega)}$ in the right-hand side of (3.9) *in terms of the data* $\boldsymbol{f}$. If we replace it by (3.8),then $\boldsymbol{f}$ is replaced by $\boldsymbol{z}$, and the resulting loss of optimality is devastating. In particular, if this is applied to the steady problem in 2-D, then we no longer know how to prove existence of solutions without restricting the data. And, much worse, we do not know how to do the numerical analysis of discrete schemes. ∎

Constructing a solution by making use of (3.1), (3.7) and (3.9) is very difficult because these three equations are redundant and no fixed-point can use all three at the same time. The originality and power of the construction by Cioranescu and Ouazar lie in that they did use all three equations. Their idea consists in discretizing (3.1) by a Galerkin method in the basis of the eigenfunctions of the operator $\mathbf{curl}\,\mathbf{curl}(\boldsymbol{u}-\alpha\,\Delta\,\boldsymbol{u})$, i.e. the functions $\boldsymbol{w}_i \in H^3(\Omega)^3 \cap V$ such that,

$$\forall \boldsymbol{v} \in H^3(\Omega)^3 \cap V\,,\ (\mathbf{curl}(\boldsymbol{w}_i-\alpha\,\Delta\,\boldsymbol{w}_i),\mathbf{curl}(\boldsymbol{v}-\alpha\,\Delta\,\boldsymbol{v})) = \lambda_i\big((\boldsymbol{w}_i,\boldsymbol{v})+\alpha\,(\nabla\,\boldsymbol{w}_i,\nabla\,\boldsymbol{v})\big)\,. \tag{3.12}$$

This special basis has the effect that, on multiplying the i-th equation that discretizes (3.1) by the eigenvalue λ_i and on summing over i, we derive a discrete version of the transport equation (3.9). This allows to recover (3.10) in the discrete case. Thus, we construct a discrete solution $\boldsymbol{u}_m$ that is bounded uniformly in $L^\infty(0,T^*;H^3(\Omega)^3)$ with $\boldsymbol{u}'_m$ bounded in $L^2(0,T^*;H^1(\Omega)^3)$. Note that all the above steps (which were hitherto formal), and in particular the delicate Green's formula (3.11), are justified because the basis functions are sufficiently smooth. Furthermore, passing to the limit is standard, because this limit is only taken in the discrete version of (3.1).

This proves local existence in time of a solution. But global existence for small data can also be established, by taking better advantage of the small damping effect of the viscous term $-\mu\Delta\,\boldsymbol{u}$. Unfortunately, Cioranescu and Ouazar did not do this in 1981 and the authors that revisited the problem from 1993 on, for example [26] or [27], did not understand the subtlety of the special basis and did not realize that, even if the 1981 results were not optimal, the method itself was optimal.

In reference [19], Cioranescu and Girault show how global existence can be achieved by the method of Cioranescu and Ouazar. The idea is to derive slightly sharper a priori estimates:

$$\|\boldsymbol{u}(t)\|^2_{L^2(\Omega)}+\alpha\,|\boldsymbol{u}(t)|^2_{H^1(\Omega)} \le \mathrm{e}^{-\mu\,K\,t}(\|\boldsymbol{u}_0\|^2_{L^2(\Omega)}+\alpha\,|\boldsymbol{u}_0|^2_{H^1(\Omega)})+\frac{S_2^2}{\mu}\int_0^t \mathrm{e}^{-\mu\,K(t-s)}\|\boldsymbol{f}(s)\|^2_{L^2(\Omega)}ds\,, \tag{3.13}$$

where

$$K=\frac{1}{\alpha+S_2^2}\,,$$

and setting $y(t)=\|\boldsymbol{z}(t)\|^2_{L^2(\Omega)}$,

$$\begin{aligned} y'(t)+\frac{\mu}{\alpha}y(t)-2\,C_2(\alpha)\,y^{3/2}(t) \le{}& \frac{4\,\mu}{\alpha^2}\mathrm{e}^{-\mu\,K\,t}(\|\boldsymbol{u}_0\|^2_{L^2(\Omega)}+\alpha\,|\boldsymbol{u}_0|^2_{H^1(\Omega)}) \\ &+\frac{4S_2^2}{\alpha^2}\int_0^t \mathrm{e}^{-\mu\,K(t-s)}\|\boldsymbol{f}(s)\|^2_{L^2(\Omega)}ds+\frac{2\,\alpha}{\mu}\|\mathbf{curl}\,\boldsymbol{f}(t)\|^2_{L^2(\Omega)}\,, \end{aligned} \tag{3.14}$$

which is indeed of Ricatti type, with the damping term $\frac{\mu}{\alpha}y(t)$ ($C(\alpha)$ is a constant that depends only on α). Owing to this damping term, we show that $y(t)$ stays bounded in $\mathbb{R}^+$ provided the data are small. This allows one to prove global existence in time of solutions, with values in $H^3(\Omega)^3$.

Regarding the regularity hypotheses on the data, it follows from (3.14) that $\mathbf{curl}\,\boldsymbol{f} \in L^2(\Omega)^3$ is sufficient (instead of $\boldsymbol{f}$ in $H^1(\Omega)^3$). As far as the domain is concerned, Bernard proves in [6] and [7] that we can take $\partial\Omega$ of class $\mathcal{C}^{2,1}$ and Ω multiply-connected. Furthermore, finding $\boldsymbol{u}$ in $H^3(\Omega)^3$ is not necessary; if we accept solutions that are less smooth, we can lower the regularity of $\partial\Omega$. Indeed, (3.14) only requires $\boldsymbol{u}$ in $W^{1,\infty}(\Omega)^3$. Thus applying Sobolev's imbedding, it suffices that $\boldsymbol{u} \in W^{2,s}(\Omega)^3$ with $s>3$. This is also sufficient for estimating $\|\boldsymbol{u}'(t)\|_{L^2(\Omega)}$.

Without details, let us describe two approaches proposed since 1993. The first one, presented in [26] consists in the Helmholtz decomposition of $\boldsymbol{u} - \alpha\,\Delta\,\boldsymbol{u}$:

$$\boldsymbol{u} - \alpha\,\Delta\,\boldsymbol{u} = \boldsymbol{w} - \nabla\, q \text{ in } \Omega\times]0,T[\,,$$

$$\operatorname{div}\boldsymbol{w} = 0 \text{ in } \Omega\times]0,T[\quad,\quad \boldsymbol{w}\cdot\boldsymbol{n} = 0 \text{ on } \partial\Omega\times]0,T[\,,$$

$$\boldsymbol{w}(0) = \boldsymbol{w}_0 \text{ on } \Omega\ ,\ \text{where } \boldsymbol{w}_0 \text{ is the Helmholtz decomposition of } \boldsymbol{u}_0\,,$$

substituted into (2.4). This gives an "Euler"-type equation involving both $\boldsymbol{w}$ and $\boldsymbol{u}$:

$$\alpha\,\frac{\partial}{\partial t}\boldsymbol{w} + \mu\,\boldsymbol{w} + \alpha\,\boldsymbol{u}\cdot\nabla\,\boldsymbol{w} + \alpha\,(\nabla\,\boldsymbol{u})^t\boldsymbol{w} - \mu\,\boldsymbol{u} + \nabla\,\hat{p} = \alpha\,\boldsymbol{f} \text{ in } \Omega\times(0,T)\,, \tag{3.15}$$

where $\hat{p}$ is another modified pressure. To prove existence of solutions, the authors solve this coupled system by Schauder's fixed point. For small data, they find $\boldsymbol{w}$ in $H^3(\Omega)^3$. But, by introducing (3.15), they lose the original equation and hence they lose the energy equality (3.6). Moreover, looking for $\boldsymbol{w}$ in $H^3(\Omega)^3$ (achieved by differentiating three times (3.15)) brings even more restrictions on the size of the data because each differentiation doubles the number of non-linear terms. For this reason, one should avoid differentiating these equations. In particular, to prove regularity of the solution of the evolution problem, one should not proceed as in [26] or [27], where existence and regularity are established in H^m, for arbitrarily large m. In the case of local existence, the interval of existence in time decreases exponentially with m, and in the case of global existence, the size of admissible data decreases also exponentially with m. This is of course unnecessary.

There is however one situation where (3.15) is useful, and that is when the curl of $\boldsymbol{f}$ is not in $L^2(\Omega)^3$. Indeed, (3.15) avoids taking the curl of (3.1). Bresch and Lemoine have used this idea in a series of publications, such as [12], to enable them to take $\boldsymbol{f}$ in $L^p(\Omega)^3$ with $p > 3$. This is more delicate than in [26], since $\boldsymbol{w}$ is no longer in $H^3(\Omega)^3$. The works of Bresch and Lemoine complete our results, but do not extend them; indeed, they cannot recover our results when $\mathbf{curl}\,\boldsymbol{f} \in L^2(\Omega)^3$, since they lose (3.6).

Remark 3.4 So far, nobody has established existence of solutions when $\boldsymbol{f} \in L^p(\Omega)^3$ with $p \le 3$; in particular, $p = 2$ is an open question. ∎

Remark 3.5 We can let α tend to zero in (3.6), but this does not appear possible in (3.14) and neither in (3.15). Nevertheless, for each given $\alpha > 0$, the other data can be adjusted so that a global solution in time exists. ∎

The second approach, analyzed by Videman [63] consists in reverting to the momentum equation (2.3) and writing $\boldsymbol{f}$ as a divergence:

$$\boldsymbol{f} = \operatorname{div}\boldsymbol{F}\,.$$

To simplify, consider the steady problem, as the process is different for the evolution problem. By introducing another modified pressure π and a tensor $\boldsymbol{\sigma}$, the equations read

$$-\Delta\,\boldsymbol{u} + \nabla\,\pi = \mathbf{div}\,\boldsymbol{\sigma}\quad,\quad \operatorname{div}\boldsymbol{u} = 0\,, \tag{3.16}$$

$$\boldsymbol{u} = \boldsymbol{0}\ \text{ on } \partial\Omega\,, \tag{3.17}$$

$$\mu\,\boldsymbol{\sigma} + \alpha\,\boldsymbol{u}\cdot\nabla\,\boldsymbol{\sigma} - \alpha\,\boldsymbol{\sigma}\cdot(\nabla\,\boldsymbol{u})^t = \boldsymbol{F} - \alpha\,\pi(\nabla\,\boldsymbol{u})^t + \alpha(\nabla\,\boldsymbol{u})^t(\nabla\,\boldsymbol{u} + (\nabla\,\boldsymbol{u})^t) - \boldsymbol{u}\otimes\boldsymbol{u}\,. \tag{3.18}$$

Again, this does not take the curl of $\boldsymbol{f}$. In [63], Videman obtains an existence result that is comparable to that of [12], i.e. for $\boldsymbol{f} \in L^p(\Omega)^3$ with $p > 3$.

3.1 The steady problem in two dimensions

The transport equation (3.9) substantially simplifies in 2-D, because $\boldsymbol{u}$ has the form $\boldsymbol{u} = (u_1, u_2, 0)$ and its curl is $\mathbf{curl}\,\boldsymbol{u} = (0, 0, \operatorname{curl}\boldsymbol{u})$ where

$$\operatorname{curl}\boldsymbol{u} = \frac{\partial u_2}{\partial x_1} - \frac{\partial u_1}{\partial x_2}\,.$$

As a consequence, $\boldsymbol{z}$ has the form $\boldsymbol{z} = (0, 0, z)$ with $z = \operatorname{curl}(\boldsymbol{u} - \alpha\Delta\,\boldsymbol{u})$, so that $\boldsymbol{z}\cdot\nabla\,\boldsymbol{u} = \mathbf{0}$. Hence (3.9) becomes a scalar equation in z:

$$\alpha\frac{\partial}{\partial t}z + \mu\,z + \alpha\,\boldsymbol{u}\cdot\nabla\,z = \mu\operatorname{curl}\boldsymbol{u} + \alpha\operatorname{curl}\boldsymbol{f} \quad \text{in } \Omega\times]0,T[\,.$$

It is no longer necessary to bound the gradient of $\boldsymbol{u}$ in $L^\infty(\Omega)$, owing that $\boldsymbol{z}\cdot\nabla\,\boldsymbol{u}$ disappears. This has allowed Ouazar to prove in [48] that, if $\partial\Omega$ is sufficiently smooth, the domain simply-connected and the boundary data zero, but without restricting the size of the other data, then the evolution problem has a unique global solution in time and the steady problem has always at least one solution.

The object of [30] is to extend this result to the steady problem with a tangential Dirichlet boundary condition, in a Lipschitz domain, possibly multiply-connected. More precisely, we show that the problem:

$$-\mu\,\Delta\,\boldsymbol{u} + \mathbf{curl}(\boldsymbol{u} - \alpha\,\Delta\,\boldsymbol{u})\times\boldsymbol{u} + \nabla\,p = \boldsymbol{f}\ ,\ \operatorname{div}\boldsymbol{u} = 0\,,$$

$$\boldsymbol{u} = \boldsymbol{g} \quad \text{on } \partial\Omega \quad \text{with} \quad \boldsymbol{g}\cdot\boldsymbol{n} = 0\,,$$

has at least one solution whatever $\boldsymbol{f} \in H(\operatorname{curl};\Omega)$, $\boldsymbol{g} \in H^{1/2}(\partial\Omega)^2$ (with $\boldsymbol{g}\cdot\boldsymbol{n} = 0$), $\mu > 0$ and $\alpha \in I\!R$. (From the mathematical point of view, here we can take $\alpha < 0$, but it may be that the problem has no physical meaning). Of course, the method of Cioranescu and Ouazar can be used to prove existence of solutions. But, since we propose to derive existence of discrete solutions further on, and as the eigenfunctions $\boldsymbol{w}_i$ do not lend themselves readily to discretization, we have chosen the following equivalent formulation, that seems better adapted to numerics. It consists in a generalized Stokes equation coupled with a transport equation, both of them linear; we denote it by *Problem P*:

• *Problem P: Find* $(\boldsymbol{u}, p, z)$ *in* $H^1_T(\Omega)\times L^2_0(\Omega)\times L^2(\Omega)$ *solution of*

$$-\mu\,\Delta\,\boldsymbol{u} + \boldsymbol{z}\times\boldsymbol{u} + \nabla\,p = \boldsymbol{f} \quad \text{with } \boldsymbol{z}\times\boldsymbol{u} = (-zu_2, zu_1)\,, \tag{3.19}$$

$$\operatorname{div}\boldsymbol{u} = 0\,, \tag{3.20}$$

$$\boldsymbol{u} = \boldsymbol{g} \quad \text{on } \partial\Omega \quad \text{with} \quad \boldsymbol{g}\cdot\boldsymbol{n} = 0\,, \tag{3.21}$$

$$\mu\,z + \alpha\,\boldsymbol{u}\cdot\nabla\,z = \mu\operatorname{curl}\boldsymbol{u} + \alpha\operatorname{curl}\boldsymbol{f}\,. \tag{3.22}$$

The crucial point here is that we prove that all solutions of (3.19)–(3.22) satisfy two energy inequalities, one that bounds $\boldsymbol{u}$ in $H^1(\Omega)^2$ and the other that bounds z in $L^2(\Omega)$, without restriction on the data. Moreover, we prove that when α tends to zero, each solution of (3.19)–(3.22) converges to a solution of the Navier-Stokes equations.

To establish the energy inequalities we need:

• a Leray–Hopf lifting of the boundary datum $\boldsymbol{g}$;

• an extension of Green's formula (3.11) to the case of a Lipschitz boundary and a velocity in H^1_T.

Let us first look at Green's formula. It is used to eliminate $\alpha(\boldsymbol{u}\cdot\nabla\,z)$ in (3.22), and it is valid in dimension n. As the right-hand side of (3.22) belongs to $L^2(\Omega)$, we see that (3.22) is a particular case of the scalar steady transport equation, in a Lipschitz domain of $I\!R^n$: Find z in $L^2(\Omega)$, such that

$$z + \mathcal{W}\,\boldsymbol{u}\cdot\nabla\,z = h \quad \text{in } \Omega\,, \tag{3.23}$$

where $\boldsymbol{u}$ is given in W, h is given in $L^2(\Omega)$ and $\mathcal{W} \in I\!R$ is a given parameter. As z and h belong to $L^2(\Omega)$, (3.23) implies that z is slightly more regular and belongs to:

$$X_{\boldsymbol{u}} = \{z \in L^2(\Omega)\,;\, \boldsymbol{u}\cdot\nabla\,z \in L^2(\Omega)\}\,, \tag{3.24}$$

for $\boldsymbol{u}$ given in W; it is a Hilbert space for the norm

$$\|z\|_{\boldsymbol{u}} = \Big(\|z\|^2_{L^2(\Omega)} + \|\boldsymbol{u} \cdot \nabla\, z\|^2_{L^2(\Omega)} \Big)^{1/2} .$$

Thus, we want to prove that

$$\forall \boldsymbol{u} \in W\,,\ \forall z \in X_{\boldsymbol{u}}\ ,\ \sum_{i=1}^{n} \int_\Omega u_i \frac{\partial z}{\partial x_i} z\, d\boldsymbol{x} = 0\,. \tag{3.25}$$

If it were known that $H^1(\Omega)$ is dense in $X_{\boldsymbol{u}}$, then (3.25) would stem trivially by density. Unfortunately, when $\boldsymbol{u}$ belongs only to H^1, this density must be established, and this is just as difficult as Green's formula itself. In fact, we shall see that these two properties are equivalent, because we shall deduce this density from Green's formula. The proof proceed in three steps:

i) First we show that the functions of $\mathcal{D}(\overline{\Omega})$ are dense in

$$\{ z \in L^2(\Omega)\,;\, \boldsymbol{u} \cdot \nabla\, z \in L^1(\Omega) \}\,,$$

for $\boldsymbol{u}$ given in $H^1(\Omega)^n$. The most difficult point is the regularization of functions in this space. The classical approach that consists first in extending functions ouside Ω is not appropriate here because it does not preserve the operation $\boldsymbol{u} \cdot \nabla\cdot$. Instead, we regularize functions by convolution with a family of mollifiers, parametrized by a set of directions and solid angles, that force the result to stay in the domain. The normal direction, the most natural choice, is not suitable, because the boundary is only Lipschitz and the normal vector is not smooth. Instead of the normal, we use the fact that the domain has the uniform cone property (equivalent to a Lipschitz condition), and we use the direction of the cone axis and its solid angle in each local chart covering the domain near the boundary. The idea of a fixed direction in each local chart is inspired by the work of Puel and Roptin [53], who use the segment property.

ii) Next, for $\boldsymbol{u} \in W$, we prove uniqueness of the solution of (3.23) in $L^1(\Omega)$. For this, we use the renormalizing of DiPerna and Lions [21]; existence of z is trivial as well as the estimates:

$$\|z\|_{L^2(\Omega)} \le \|h\|_{L^2(\Omega)}\ ,\ |\mathcal{W}| \|\boldsymbol{u} \cdot \nabla\, z\|_{L^2(\Omega)} \le \|h\|_{L^2(\Omega)}\,. \tag{3.26}$$

iii) Finally, we establish Green's formula (3.25) and the density of $\mathcal{D}(\Omega)$ in $X_{\boldsymbol{u}}$.

Remark 3.6 The density in i) holds without restriction on $\boldsymbol{u}$. But is $\mathcal{D}(\overline{\Omega})$ dense in $X_{\boldsymbol{u}}$ when $\boldsymbol{u}$ is arbitrary in $H^1(\Omega)^n$? If we knew this were true, we could give meaning to the left-hand side of (3.11) and we could solve the steady problem with any Dirichlet boundary condition, by imposing $\boldsymbol{z}$ where $\boldsymbol{u} \cdot \boldsymbol{n} \neq 0$. ∎

The first estimate in (3.26) can be generalized to exponents $p > 2$. If $h \in L^p(\Omega)$ with $p > 2$, by extending a result due to Ortega [47] (written on a smooth domain with a driving velocity $\boldsymbol{u} \in W^{1,\infty}(\Omega)^n \cap H^1_0(\Omega)^n$), we easily prove that the solution z of (3.23), for $\boldsymbol{u} \in W$ and Ω Lipschitz, belongs to $L^p(\Omega)$ and

$$\|z\|_{L^p(\Omega)} \le \|h\|_{L^p(\Omega)}\,. \tag{3.27}$$

If $p < 2$, with $p > 1$ when $n = 2$ and $p > 2n/(n+2)$ when $n \ge 3$, by proceeding by duality and transposition (cf. [45]), we show that the transposed equation has one and only one solution $z \in L^p(\Omega)$ that satisfies (3.27) and that solves (3.23). Then, a fixed-point argument on $\boldsymbol{u}$ shows that in two dimensions, *Problem P* has at least one solution when $\boldsymbol{f} \in L^2(\Omega)^2$ with $\operatorname{curl} \boldsymbol{f} \in L^p(\Omega)$ for $p > 1$ and $\boldsymbol{g} = \boldsymbol{0}$.

Remark 3.7 There remain many open questions concerning (3.23). For instance, what are minimal conditions on Ω for z to belong to $H^1(\Omega)$? A formal differentiation of (3.23) yields the sufficient condition on h and $\boldsymbol{u}$: $h \in H^1(\Omega)$ and $\boldsymbol{u} \in W \cap W^{1,\infty}(\Omega)^n$, small enough. More precisely, we obtain formally:

$$\|\nabla\, z\|_{L^2(\Omega)} \big(1 - |\mathcal{W}| \|\nabla\, \boldsymbol{u}\|_{L^\infty(\Omega)} \big) \le \|\nabla\, h\|_{L^2(\Omega)}\,.$$

But we do not know how to justify this inequality, without asking either $\partial\Omega$ smooth or Ω convex (in 2 or 3-D), because we use the regularity of a Laplace equation. This brings us to another question: what are minimal conditions on Ω and $\boldsymbol{u}$ for $z \in H^\theta(\Omega)$, with $\theta \in]0, 1/2]$? ∎

Now we turn to the Leray–Hopf's lifting. We need it to show existence of solutions of the generalized Stokes system (3.19)–(3.21). This problem has the equivalent variational formulation: Find $\boldsymbol{u} \in V + \boldsymbol{w_g}$ such that

$$\forall \boldsymbol{v} \in V\,,\ \mu(\nabla\,\boldsymbol{u}, \nabla\,\boldsymbol{v}) + (\boldsymbol{z} \times \boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{f}, \boldsymbol{v})\,,$$

where $\boldsymbol{w_g}$ is a lifting of $\boldsymbol{g}$ in W (hence divergence-free). It is well-known that several liftings exist. At first sight, the most natural choice is the solution of a non-homogeneous Stokes problem:

$$-\Delta\,\boldsymbol{w_g} + \nabla\,p_{\boldsymbol{g}} = \boldsymbol{0} \ \text{ and } \ \operatorname{div}\boldsymbol{w_g} = 0 \ \text{ in } \Omega \quad , \quad \boldsymbol{w_g} = \boldsymbol{g} \text{ on } \partial\Omega\,. \tag{3.28}$$

This problem has a unique solution that depends continuously on $\boldsymbol{g}$: there exists a constant $\mathcal{L}$ such that (cf. for instance [29]):

$$|\boldsymbol{w_g}|_{H^1(\Omega)} \le \mathcal{L}\|\boldsymbol{g}\|_{H^{1/2}(\partial\Omega)}\,. \tag{3.29}$$

Now, for estimating $\boldsymbol{u}$, we consider the equation satisfied by $\boldsymbol{u}_0 = \boldsymbol{u} - \boldsymbol{w_g}$:

$$\forall \boldsymbol{v} \in V\,,\ \mu(\nabla\,\boldsymbol{u}_0, \nabla\,\boldsymbol{v}) + (\boldsymbol{z} \times \boldsymbol{u}_0, \boldsymbol{v}) = (\boldsymbol{f}, \boldsymbol{v}) - \mu(\nabla\,\boldsymbol{w_g}, \nabla\,\boldsymbol{v}) - (\boldsymbol{z} \times \boldsymbol{w_g}, \boldsymbol{v})\,,$$

that simplifies because $(\nabla\,\boldsymbol{w_g}, \nabla\,\boldsymbol{v}) = 0$ and yields the estimate:

$$\begin{aligned} |\boldsymbol{u}_0|_{H^1(\Omega)} &\le \frac{1}{\mu}\big(S_2\|\boldsymbol{f}\|_{L^2(\Omega)} + \|\boldsymbol{z}\|_{L^2(\Omega)} \sup_{\boldsymbol{v}\in V} \frac{\||\boldsymbol{w_g}|\,|\boldsymbol{v}|\|_{L^2(\Omega)}}{|\boldsymbol{v}|_{H^1(\Omega)}}\big) \\ &\le \frac{1}{\mu}\big(S_2\|\boldsymbol{f}\|_{L^2(\Omega)} + S_4\tilde{S}_4\|\boldsymbol{z}\|_{L^2(\Omega)}|\boldsymbol{w_g}|_{H^1(\Omega)}\big)\,. \end{aligned} \tag{3.30}$$

But this last estimate is usually not sufficiently sharp because it involves $\|\boldsymbol{z}\|_{L^2(\Omega)}$ multiplied by a factor that is not necessarily small, except when $\boldsymbol{g}$ is small enough or μ large enough. Hence this $\boldsymbol{w_g}$ is not always convenient. A closer scrutiny at the first part of (3.30) reveals that it would be desirable to bound $\||\boldsymbol{w_g}|\,|\boldsymbol{v}|\|_{L^2(\Omega)}$ by $\varepsilon\,|\boldsymbol{v}|_{H^1(\Omega)}$ for arbitrary ε; this property is typical of Leray–Hopf's lifting (cf. [39], [42]).

The idea for constructing this lifting (cf. [44] or [29]) consists in truncating $\boldsymbol{w_g}$ so that it is supported by a small neighborhood of the boundary, with "width" related to the parameter ε. But since truncation does not preserve the zero divergence, the stream function of $\boldsymbol{w_g}$ (or vector potential in 3-D) is truncated; this stream function exists because $\boldsymbol{g}\cdot\boldsymbol{n} = 0$. The disadvantage of the classical construction in [44] and [29] is that the "width" of this support tends to zero exponentially with μ. From a theoretical point of view, this is unimportant. From the approximation point of view, this is a serious drawback when μ is small (even though the lifting is never computed), because it means that in order to prove existence of a discrete solution, we must use a very fine mesh (possibly unrealistic) near the boundary.

However, in the case where $\boldsymbol{g}\cdot\boldsymbol{n} = 0$, this specific truncation is not necessary. Reference [32] constructs a lifting $\boldsymbol{u_g}$ supported in a neighborhood whose "width" is of the order of μ, when Ω is Lipschitz and $\boldsymbol{g}$ belongs either to $W^{1-1/\lambda,\lambda}(\partial\Omega)^2$ for some $\lambda > 2$, or to $H^{1/2}(\partial\Omega)^2$ when the boundary is a polygon. This extends a result of [61], proven when $\partial\Omega$ is of class $\mathcal{C}^3$ and $\boldsymbol{g} \in \mathcal{C}^3(\partial\Omega)^2$. When $\boldsymbol{g} \in H^{1/2}(\partial\Omega)^2$ and Ω is a Lipschitz polygon (i.e. without cracks), we find (cf. [32]):

$$\|\nabla\,\boldsymbol{u_g}\|_{L^2(\Omega_\varepsilon)} \le C\frac{1}{\varepsilon^{1/2+1/\tau}}\|\boldsymbol{g}\|_{H^{1/2}(\partial\Omega)}\ ,\ \text{for } 2 \le \tau < \infty\,, \tag{3.31}$$

$$\forall \boldsymbol{v} \in H^1_0(\Omega)^2\,,\ \||\boldsymbol{u_g}|\,|\boldsymbol{v}|\|_{L^2(\Omega_\varepsilon)} \le C\varepsilon^{1/\tau}|\boldsymbol{v}|_{H^1(\Omega_\varepsilon)}\|\boldsymbol{g}\|_{H^{1/2}(\partial\Omega)}\ ,\ \text{for } 1 < \tau\,, \tag{3.32}$$

where the constants C depend τ, but neither on $\boldsymbol{g}$ nor on ε. Thus, we obtain for each number $\tau > \frac{1}{2}$:

$$\|z\|_{L^2(\Omega)} \le 2\frac{|\alpha|}{\mu}\|\operatorname{curl}\boldsymbol{f}\|_{L^2(\Omega)} + 2\frac{\sqrt{2}}{\mu}S_2\|\boldsymbol{f}\|_{L^2(\Omega)} + \frac{C}{\mu^\tau}\|\boldsymbol{g}\|^{1+\tau}_{H^{1/2}(\partial\Omega)}\,, \tag{3.33}$$

where C depends only on τ and Ω. When Ω is an arbitrary Lipschitz domain and $\boldsymbol{g} \in W^{1-1/\lambda,\lambda}(\partial\Omega)^2$, for some $\lambda > 2$, then there exists a constant C depending only on λ and Ω, such that:

$$\|z\|_{L^2(\Omega)} \le 2\frac{|\alpha|}{\mu}\|\operatorname{curl}\boldsymbol{f}\|_{L^2(\Omega)} + 2\frac{\sqrt{2}}{\mu}S_2\|\boldsymbol{f}\|_{L^2(\Omega)} + (2\sqrt{2})^{3/2}\frac{C}{\mu^{1/2}}\|\boldsymbol{g}\|^{3/2}_{W^{1-1/\lambda,\lambda}(\partial\Omega)}\,. \tag{3.34}$$

The factor $1/\mu$ in the first two terms of (3.33) and (3.34) is inevitable: it arises even in the homogeneous case. In contrast, the factor $1/\sqrt{\mu}$ comes from the non-homogeneous boundary term and is negligible with respect to an exponential. Summing up, when μ is small, we have gained substantially by not using the classical lifting.

Now we prove existence of solutions of *Problem P*. Let $\{w_i\}_{i\geq 1}$ be a basis of $H^2(\Omega)$ and Z_m the vector space spanned by w_i for $1 \leq i \leq m$. We discretize z by Galerkin's method in this basis. For each $z_m \in Z_m$, we set $\boldsymbol{z}_m = (0, 0, z_m)$ and we note $\boldsymbol{u}(z_m), p(z_m)$ the unique solution of the generalized Stokes problem (3.19)–(3.21) with z_m instead of z. Next, we discretize the transport equation (3.22) by: Find z_m in Z_m solution of, for $1 \leq i \leq m$,

$$\mu\,(z_m, w_i) + \alpha\,(\boldsymbol{u}(z_m)\cdot\nabla\, z_m, w_i) = \mu(\operatorname{curl}\boldsymbol{u}(z_m), w_i) + \alpha\,(\operatorname{curl}\boldsymbol{f}, w_i)\,. \tag{3.35}$$

Observe that $\boldsymbol{u}(z_m)$ belongs to a finite-dimensional space because so does z_m. Hence, we can apply here Brouwer's fixed-point theorem. It implies existence of a solution z_m satisfying the uniform estimates in m (3.33) or (3.34). These estimates allow one to pass to the limit in (3.35) and (3.19)–(3.21), whence existence of a solution of *Problem P* in $H^1_T(\Omega)\times L^2_0(\Omega)\times L^2(\Omega)$, without restriction on the data. Finally, Green's formula (3.25) shows that all solutions of this problem satisfy these estimates.

These solutions depend on μ and α. In the estimates (3.33) and (3.34), we cannot let μ tend to zero, but we can let α tend to zero. Green's formula (3.25) and a similar limiting process allow one to prove that when α tends to zero, each solution of *Problem P* converges strongly in $H^1_T(\Omega)\times L^2_0(\Omega)\times L^2(\Omega)$ to a solution of the steady incompressible Navier-Stokes equations.

Remark 3.8 We have seen that (3.19)–(3.21) define $\boldsymbol{u}$ in $H^1_T(\Omega)$ in terms of z in $L^2(\Omega)$. Moreover, $\boldsymbol{u}$ is locally Lipschitz with respect to z. Indeed, set $\boldsymbol{u}_1 = \boldsymbol{u}(z)$ and $\boldsymbol{u}_2 = \boldsymbol{u}(z+\zeta)$ for arbitrary z and ζ in $L^2(\Omega)$; we have

$$|\boldsymbol{u}_1 - \boldsymbol{u}_2|_{H^1(\Omega)} \leq \frac{S_4\tilde{S}_4}{\mu}|\boldsymbol{u}_2|_{H^1(\Omega)}\|\zeta\|_{L^2(\Omega)}\,. \tag{3.36}$$

We have also seen that (3.22) defines z in $L^2(\Omega)$ in terms of $\boldsymbol{u}$ in $H^1_T(\Omega)$, but in these spaces z *is not locally Lipschitz* with respect to $\boldsymbol{u}$. Indeed, let z_1 be the solution of (3.22) associated with $\boldsymbol{u}$ and z_2 the solution of (3.22) associated with $\boldsymbol{u}+\boldsymbol{v}$ where $\boldsymbol{u}$ and $\boldsymbol{v}$ are arbitrary in $H^1_T(\Omega)$. Then,

$$\mu\,(z_2 - z_1) + \alpha\,(\boldsymbol{u}+\boldsymbol{v})\cdot\nabla(z_2 - z_1) = \mu\operatorname{curl}\boldsymbol{v} - \alpha\,\boldsymbol{v}\,\cdot\nabla\, z_1\,,$$

and we do not know how to bound the last term, because there is no reason why $\alpha\,\boldsymbol{v}\cdot\nabla\, z_1$ should be in $L^2(\Omega)$. For this term to be in $L^2(\Omega)$, it suffices to ask for example that $\boldsymbol{v}\in L^\infty(\Omega)^2$ and $z_1 \in H^1(\Omega)$. Now, a sufficient condition for $z_1 \in H^1(\Omega)$ is (cf. Remark 3.7): Ω convex, $\boldsymbol{u}\in (W^{1,\infty}(\Omega)\cap H^2(\Omega))^2$, $\operatorname{curl}\boldsymbol{f}\in H^1(\Omega)$ and

$$|\alpha|\|\nabla\,\boldsymbol{u}\|_{L^\infty(\Omega)} \leq \mu - \eta \;\text{ for a number } \eta > 0, \eta < \mu\,. \tag{3.37}$$

In this case, we have

$$\|z_2 - z_1\|_{L^2(\Omega)} \leq \|\operatorname{curl}\boldsymbol{v}\|_{L^2(\Omega)} + \frac{|\alpha|}{\mu}\|\boldsymbol{v}\|_{L^\infty(\Omega)}|z_1|_{H^1(\Omega)}\,.$$

The same argument shows that the mapping $\boldsymbol{u}\mapsto z$ from $H^1_T(\Omega)$ with values in $X_{\boldsymbol{u}}$ is not differentiable. This explains the poor performance of the Implicit Function Theorem when applied to the discrete *P*. ∎

4 Approximation in two dimensions

From now on, we assume that the domain Ω is a Lipschitz polygon, so it can be entirely triangulated. We shall discuss here only two discrete schemes for *Problem P*: one that uses a centered approximation of the transport term and one that uses an upwind approximation of the discontinuous Galerkin type. The

reader can refer to [31] and [33] for other schemes. The numerical analysis of the upwind approximation is more technical than that of the centered approximation, but it usually gives better results.

The crucial point of the preceding section is that *Problem P* has at least one solution, without restriction on the data (exactly as for the incompressible steady Navier-Stokes system) and for constructing a solution it is not necessary that the velocity gradient be bounded in L^∞. The schemes presented here are chosen so that this bound is not necessary in the discrete case. This will enable us to really do their numerical analysis; otherwise, this analysis is an open problem. Thus, the finite element spaces are chosen to satisfy three criteria:
• the schemes must have a solution in a Lipschitz polygon, without restriction on the data,
• always without restriction, each discrete solution must converge strongly to a solution of the exact problem, as the mesh-size tends to zero,
• under suitable restrictions on the data and the angles of the polygon, the discrete solutions must satisfy error inequalities that lead to error estimates.

Let us triangulate the domain. Let κ be an arbitrary triangle; we note h_κ its diameter and ρ_κ the diameter of its inscribed circle. Let $h > 0$ be a discretization parameter and $\mathcal{T}_h$ a *regular* family of triangulations of $\overline{\Omega}$, made of triangles with maximum diameter h, i.e.

$$h := \max_{\kappa\in\mathcal{T}_h} h_\kappa \ , \ \max_{\kappa\in\mathcal{T}_h} \frac{h_\kappa}{\rho_\kappa} \leq \sigma_0 \, , \tag{4.1}$$

with a constant σ_0 independent of h (cf. [16]). As usual, the triangulation is such that any pair of triangles are either disjoint, or share a vertex, or a complete side.

4.1 A centered approximation

We first describe a general centered approximation with continuous pressure. Let $X_{h,T}$ be a finite-element space made of continuous vector-valued functions, with vanishing tangential trace on $\partial\Omega$, $X_h = X_{h,T} \cap H^1_0(\Omega)^2$, and M_h and Z_h be two finite-element spaces made also of continuous functions, the functions of M_h having *zero mean-value*. We suppose that $\boldsymbol{g}_h$ is a suitable approximation of $\boldsymbol{g}$ extended to Ω, specified further on. We approximate *Problem P* by: *Find* $\boldsymbol{u}_h$ *in* $X_h+\boldsymbol{g}_h$, p_h *in* M_h *and* $\boldsymbol{z}_h = (0,0,z_h)$ *with* z_h *in* Z_h, *such that*

$$\forall \boldsymbol{v}_h \in X_h \, , \, \mu(\nabla \boldsymbol{u}_h, \nabla \boldsymbol{v}_h) + (\boldsymbol{z}_h \times \boldsymbol{u}_h, \boldsymbol{v}_h) - (p_h, \operatorname{div} \boldsymbol{v}_h) = (\boldsymbol{f}, \boldsymbol{v}_h) \, , \tag{4.2}$$

$$\forall q_h \in M_h \, , \, (q_h, \operatorname{div} \boldsymbol{u}_h) = 0 \, , \tag{4.3}$$

$$\forall \theta_h \in Z_h \, , \, \mu \, (z_h, \theta_h) + \alpha \, (\boldsymbol{u}_h \cdot \nabla \, z_h, \theta_h) + \frac{\alpha}{2}((\operatorname{div} \boldsymbol{u}_h) z_h, \theta_h) = \mu \, (\operatorname{curl} \boldsymbol{u}_h, \theta_h) + \alpha \, (\operatorname{curl} \boldsymbol{f}, \theta_h) \, . \tag{4.4}$$

As usual, (cf. [60]), the last term in the left-hand side is chosen so that z_h satisfies:

$$\mu \, \|z_h\|^2_{L^2(\Omega)} = \mu(\operatorname{curl} \boldsymbol{u}_h, z_h) + \alpha(\operatorname{curl} \boldsymbol{f}, z_h) \, . \tag{4.5}$$

We note W_h and V_h the spaces:

$$W_h = \{\boldsymbol{v} \in X_{h,T} \, ; \, \forall q \in M_h \, , \, \int_\Omega q \operatorname{div} \boldsymbol{v} \, d\boldsymbol{x} = 0\} \ , \ V_h = W_h \cap H^1_0(\Omega)^2 \, . \tag{4.6}$$

As in all approximations of incompressible fluids, the spaces X_h and M_h are not independent. They must satisfy a compatibility condition, the discrete inf-sup condition of Babuška-Brezzi (cf. [3] or [13]), uniform with respect to h: there exists a constant $\beta^* > 0$, independent of h, such that for all $q_h \in M_h$,

$$\sup_{\boldsymbol{v}_h \in X_h} \frac{\int_\Omega q_h \operatorname{div} \boldsymbol{v}_h \, d\boldsymbol{x}}{|\boldsymbol{v}_h|_{H^1(\Omega)}} \geq \beta^* \|q_h\|_{L^2(\Omega)} \, . \tag{4.7}$$

Similarly, to prove existence of discrete solutions, $\boldsymbol{g}_h$ cannot be an arbitrary approximation of $\boldsymbol{g}$ in $X_{h,T}$; it must belong to W_h. This raises the same question on the lifting as for the exact problem: what is the

choice of $\boldsymbol{g}_h$ for $\||\boldsymbol{g}_h|\,|\boldsymbol{v}|\|_{L^2(\Omega)}$ to be bounded by $\varepsilon\,|\boldsymbol{v}|_{H^1(\Omega)}$ with ε arbitrary ? The simplest idea consists in approximating well the lifting function $\boldsymbol{u_g}$. Therefore this approximation should belong to W_h and mimick as best as possible the estimate (3.32). More precisely, suppose we have an operator P_h such that $P_h(\boldsymbol{u_g}) \in W_h$ and write

$$\||P_h(\boldsymbol{u_g})|\,|\boldsymbol{v}|\|_{L^2(\Omega)} \leq \||P_h(\boldsymbol{u_g}) - \boldsymbol{u_g}|\,|\boldsymbol{v}|\|_{L^2(\Omega)} + \||\boldsymbol{u_g}|\,|\boldsymbol{v}|\|_{L^2(\Omega)}\,.$$

If P_h satisfies an optimal error estimate in $L^{2+\gamma}$ for $\gamma > 0$:

$$\|P_h(\boldsymbol{u_g}) - \boldsymbol{u_g}\|_{L^{2+\gamma}(\Omega)} \leq Ch^{\frac{2}{2+\gamma}}|\boldsymbol{u_g}|_{H^1(\Omega)}\,,$$

then in view of (3.31), we see that if $h < \varepsilon$, then (3.32) holds with the factor $\varepsilon^{1/2-\delta}$ instead of $\varepsilon^{1-\delta}$, and hence ε is of the order of μ^2. This loss of $\varepsilon^{1/2}$ is due to the fact that the discrete velocities are not exactly divergence-free. When their divergence is zero, then the stream-function is directly approximated and we recover the factor $\varepsilon^{1-\delta}$. The corresponding finite elements are more costly, but they allow to save on the mesh-size.

The condition $h < \varepsilon$ is very restrictive when ε is small, because it imposes a very small mesh-size throughout the domain, solely due to the boundary datum. But if the support of $P_h(\boldsymbol{u_g})$ is close to that of $\boldsymbol{u_g}$, then it can be replaced by a condition on the mesh-size near the boundary, since the support of $\boldsymbol{u_g}$ is concentrated there. Thus, we need only refine the triangulation near the boundary.

Besides this, to estimate the error of (4.2)–(4.4), we shall need further on the following decoupling inequality:

$$\begin{aligned}|\boldsymbol{u}_h - \boldsymbol{u}|_{W^{1,r}(\Omega)} \leq |P_h(\boldsymbol{u}) - \boldsymbol{u}|_{W^{1,r}(\Omega)} + Ch^{2/r-1}(K_1(h)|P_h(\boldsymbol{u}) - \boldsymbol{u}|_{H^1(\Omega)} + \|p - r_h(p)\|_{L^2(\Omega)})\\ + CK_2(h)\big(1 + h^{2/r-1/2}(1 + K_1(h))\big)\|z - z_h\|_{L^2(\Omega)}\,,\end{aligned} \tag{4.8}$$

for $r \in [2, 4]$, where K_1 and K_2 are defined further on. It enables to write the error on $\boldsymbol{u}$ in terms of that on z. But to take good advantage of (4.8), we need a sharp estimate on the error of P_h in norm $W^{1,r}$.

From these considerations, it stems that we need an approximation operator that:
- preserves the discrete divergence,
- preserves approximately the support,
- and has optimal approximation properties in L^p and $W^{1,p}$ for $p > 2$.

If an explicit and local construction of this operator is known, then these properties are easily proven. But in the case of Taylor-Hood finite elements that we shall use here, this explicit construction is not known. We cannot use directly Fortin's Lemma, cf. [25] or [29], proving that the inf-sup condition (4.7) guarantees automatically existence of an approximation operator preserving the discrete divergence and stable in H^1. Indeed, this operator is based on the solution of a discrete Stokes problem in the domain. On one hand, this solution cannot (even approximately) preserve the support, because it is supported by the whole domain. On the other hand, a good approximation estimate in L^p is derived by a duality argument, which according to the values of p imposes restrictions on the angles of the boundary. Finally, proving an optimal approximation estimate in $W^{1,p}$ for $p > 2$ was until recently an open problem: it was originally proven by [24] with a logarithmic factor.

In [31] and [34], we construct a quasi-local approximation operator that has the same properties as a standard interpolation operator and that preserves the discrete divergence. The idea is to modify appropriately the proof of the inf-sup condition proposed by [10] and [58] (cf. [29]). This proof proceeds in two steps. First, a local inf-sup condition is established for discrete pressures with local zero mean-value. Next passing to arbitrary discrete pressures is achieved by proving a global inf-sup condition for piecewise constant pressures. The modification proposed here consists in eliminating this second step, because it is not local.

Now, we describe the finite element spaces (cf. [38]):

$$X_{h,T} = \{\boldsymbol{v}_h \in \mathcal{C}^0(\overline{\Omega})^2\,;\, \forall \kappa \in \mathcal{T}_h\,,\, \boldsymbol{v}_h|_\kappa \in I\!P_2^2\} \cap H^1_T(\Omega)\ ,\ X_h = X_{h,T} \cap H^1_0(\Omega)^2\,, \tag{4.9}$$

$$M_h = \{q_h \in \mathcal{C}^0(\overline{\Omega})\,;\, \forall \kappa \in \mathcal{T}_h\,,\, q_h|_\kappa \in I\!P_1\} \cap L^2_0(\Omega)\,, \tag{4.10}$$

$$Z_h = \{\theta_h \in \mathcal{C}^0(\overline{\Omega})\,;\, \forall \kappa \in \mathcal{T}_h\,,\, \theta_h|_\kappa \in I\!P_1\}\,. \tag{4.11}$$

The inf-sup condition for this element was first established by [5], next by [62] and later by [29] and by [15]. All these proofs have in common the assumption that each triangle κ has at most one side on $\partial\Omega$.

Eliminating the step with piecewise constant pressures stems from the observation that the above references construct in each element an auxiliary velocity in the space

$$\tilde{X}_h = \{\boldsymbol{v}_h \in X_h\,;\, \forall \kappa \in \mathcal{T}_h\,,\, \int_\kappa \operatorname{div} \boldsymbol{v}_h \, d\boldsymbol{x} = 0\}\,,$$

and establish a local weak inf-sup condition for each pressure in

$$\tilde{M}_h = \{\tilde{q}_h\,;\, q_h \in M_h\}\,, \text{ where } \tilde{q}_h = q_h - \frac{1}{|\kappa|}\int_\kappa q_h \, d\boldsymbol{x}\,.$$

It is weak in the sense that the associated velocities do not vanish on the boundary of κ. Thus, if $\Pi_h \in \mathcal{L}(H^1_T(\Omega); X_{h,T})$ is an auxiliary interpolation operator that satisfies for all $\boldsymbol{v} \in H^1_T(\Omega)$:

$$\forall \kappa \in \mathcal{T}_h\,,\, \int_\kappa \operatorname{div}(\Pi_h(\boldsymbol{v}) - \boldsymbol{v}) d\boldsymbol{x} = 0\,, \tag{4.12}$$

and has optimal approximation properties, this inf-sup condition allows one to define P_h by:

$$P_h(\boldsymbol{v}) = \Pi_h(\boldsymbol{v}) + \boldsymbol{c}_h(\boldsymbol{v})\,, \tag{4.13}$$

where the correction $\boldsymbol{c}_h(\boldsymbol{v}) \in \tilde{X}_h$ is the solution of

$$\forall q_h \in \tilde{M}_h\,,\, \int_\Omega q_h \operatorname{div} \boldsymbol{c}_h(\boldsymbol{v})\, d\boldsymbol{x} = \int_\Omega q_h \operatorname{div}(\boldsymbol{v} - \Pi_h(\boldsymbol{v}))\, d\boldsymbol{x}\,. \tag{4.14}$$

Owing to (4.12), (4.14) and the definition of $\tilde{X}_h$, we have that P_h preserves automatically the discrete divergence, i.e.

$$\forall \boldsymbol{w} \in H^1_T(\Omega)\,,\, \forall q_h \in M_h\,,\, \int_\Omega q_h \operatorname{div}(P_h(\boldsymbol{w}) - \boldsymbol{w})\, d\boldsymbol{x} = 0\,. \tag{4.15}$$

To guarantee the quasi-local character of P_h, the elements are grouped into "star-like" macro-elements (with or without overlaps) that share a common vertex, and $\boldsymbol{c}_h(\boldsymbol{v})$ is made to vanish on the boundary of each macro-element. Furthermore, in each macro-element, the inf-sup condition holds for any norm, since the macro-element involves only finite-dimensional spaces on which all norms are equivalent. Hence we can prove that:

$$\forall \boldsymbol{v} \in H^1_T(\Omega)\,,\, \|\boldsymbol{v} - P_h(\boldsymbol{v})\|_{L^p(\Omega)} \leq C\, h^{2/p} |\boldsymbol{v}|_{H^1(\Omega)}\,, \tag{4.16}$$

$$\forall \boldsymbol{v} \in H^1_T(\Omega) \cap W^{s,p}(\Omega)^2\,,\, |\boldsymbol{v} - P_h(\boldsymbol{v})|_{W^{m,p}(\Omega)} \leq C\, h^{s-m} |\boldsymbol{v}|_{W^{s,p}(\Omega)}\,, \tag{4.17}$$

for all p with $2 \leq p \leq \infty$ and all s with $1 \leq s \leq 3$, $m = 0, 1$. And by construction, the support of $P_h(\boldsymbol{v})$ satisfies:

$$\operatorname{dist}\big(\operatorname{supp}(P_h(\boldsymbol{v})), \operatorname{supp}(\boldsymbol{v})\big) \leq C\, h\,, \tag{4.18}$$

with a constant C independent of h.

Let us revert to the discrete problem (4.2)–(4.4). In view of the above considerations, we take $\boldsymbol{g}_h = P_h(\boldsymbol{u}_{\boldsymbol{g}})$. Note that the expression of $\boldsymbol{u}_{\boldsymbol{g}}$ is not necessary for this, because if $\boldsymbol{w}$ is another lifting of $\boldsymbol{g}$, then $\boldsymbol{u}_{\boldsymbol{g}} - \boldsymbol{w}$ vanishes on the boundary, and since P_h preserves the zero boundary values, we have

$$P_h(\boldsymbol{w})|_{\partial\Omega} = P_h(\boldsymbol{u}_{\boldsymbol{g}})|_{\partial\Omega}\,. \tag{4.19}$$

This is why, in practice, $\boldsymbol{g}_h$ is needed only on the boundary and is constructed by interpolating directly $\boldsymbol{g}$, with an interpolation that preserves the vanishing tangential trace.

Besides this, considering that $\boldsymbol{f}$ is fixed, for a given z_h, the solution $(\boldsymbol{u}_h, p_h)$ of (4.2), (4.3) depends only on the trace of $\boldsymbol{g}_h$. It is unique and does not depend on the choice of $\boldsymbol{g}_h$ inside the domain. Thus, according to this choice, we find a variety of estimates for $\boldsymbol{u}_h$ and p_h in terms of z_h. In particular, we can also use the discrete analogue $\boldsymbol{w}_h$ of $\boldsymbol{w}_{\boldsymbol{g}}$ (cf.(3.28)) defined by: $\boldsymbol{w}_h \in V_h + \boldsymbol{g}_h$ unique solution of:

$$\forall \boldsymbol{v}_h \in V_h\,,\ (\nabla\, \boldsymbol{w}_h, \nabla\, \boldsymbol{v}_h) = 0\,. \tag{4.20}$$

Observing that

$$\forall \boldsymbol{v}_h \in V_h + \boldsymbol{g}_h\ , |\boldsymbol{w}_h|_{H^1(\Omega)} \leq |\boldsymbol{v}_h|_{H^1(\Omega)}\,,$$

and taking $\boldsymbol{v}_h = P_h(\boldsymbol{w}_{\boldsymbol{g}})$, we find

$$|\boldsymbol{w}_h|_{H^1(\Omega)} \leq |P_h(\boldsymbol{w}_{\boldsymbol{g}})|_{H^1(\Omega)} \leq C_1 |\boldsymbol{w}_{\boldsymbol{g}}|_{H^1(\Omega)} \leq C_1 \mathcal{L} \|\boldsymbol{g}\|_{H^{1/2}(\partial\Omega)}\,, \tag{4.21}$$

where the constant C_1 is derived from (4.17) and $\mathcal{L}$ is the constant of (3.29). With these two liftings, it is easy to show existence of a solution of (4.2)–(4.4):

Lemma 4.1 *Under the above assumptions on the triangulation, for each $z_h \in Z_h$, (4.2), (4.3) has a unique solution $\boldsymbol{u}_h \in X_h + \boldsymbol{g}_h$. This solution satisfies:*

$$|\boldsymbol{u}_h|_{H^1(\Omega)} \leq \frac{S_2}{\mu} \|\boldsymbol{f}\|_{L^2(\Omega)} + K_1(h) C_1 \mathcal{L} \|\boldsymbol{g}\|_{H^{1/2}(\partial\Omega)}\,, \tag{4.22}$$

where

$$K_1(h) = 1 + \frac{S_4 \tilde{S}_4}{\mu} \|z_h\|_{L^2(\Omega)}\,,$$

$$\|p_h\|_{L^2(\Omega)} \leq \frac{1}{\beta^*} \big(S_2 \|\boldsymbol{f}\|_{L^2(\Omega)} + \mu C_1 \mathcal{L} \|\boldsymbol{g}\|_{H^{1/2}(\partial\Omega)} + S_4 \tilde{S}_4 |\boldsymbol{u}_h|_{H^1(\Omega)} \|z_h\|_{L^2(\Omega)} \big)\,. \tag{4.23}$$

Moreover, there exists a constant $C_2 > 0$, independent of h, such that for all $\varepsilon > 0$, if for a number $\tau > 0$,

$$h_b < C_2 \varepsilon^{2+\tau} \|\boldsymbol{g}\|^{-2-\tau}_{H^{1/2}(\partial\Omega)}\,, \tag{4.24}$$

where h_b is the maximum diameter of the elements in a neighborhood of $\partial\Omega$, then for all $s > \frac{\tau}{2}$, we have

$$|\boldsymbol{u}_h|_{H^1(\Omega)} \leq \frac{S_2}{\mu} \|\boldsymbol{f}\|_{L^2(\Omega)} + \frac{C_3}{\varepsilon^{1+s}} \|\boldsymbol{g}\|^{2+s}_{H^{1/2}(\partial\Omega)} + \frac{\varepsilon}{\mu} \|z_h\|_{L^2(\Omega)}\,, \tag{4.25}$$

where the constant C_3 depends on s and τ, but not on h, μ and ε.

On substituting (4.25) into (4.5) and choosing

$$\varepsilon = \frac{\mu}{2\sqrt{2}}\,,$$

we infer the following estimate for z_h.

Theorem 4.2 *Under the above assumptions on the triangulation, the constant C_2 of (4.24) is such that for all $\mu > 0$ and $\alpha \in \mathbb{R}$, for all $\boldsymbol{f}$ in $H(curl, \Omega)$ and all $\boldsymbol{g}$ in $H^{1/2}(\partial\Omega)^2$ satisfying $\boldsymbol{g} \cdot \boldsymbol{n} = 0$, if*

$$h_b < C_2 \big(\frac{\mu}{2\sqrt{2}}\big)^{2+\tau} \|\boldsymbol{g}\|^{-2-\tau}_{H^{1/2}(\partial\Omega)}\,,\ \ \textit{for some } \tau > 0\,, \tag{4.26}$$

then (4.2)–(4.4) has at least one solution $(\boldsymbol{u}_h, p_h, z_h)$ in $(X_h + \boldsymbol{g}_h) \times M_h \times Z_h$ and each solution satisfies the a priori estimates (4.22), (4.23) and (4.25) with the same constant C_3. In addition, for all $s > \frac{\tau}{2}$,

$$\|z_h\|_{L^2(\Omega)} \leq 2\sqrt{2} \big(\frac{S_2}{\mu} \|\boldsymbol{f}\|_{L^2(\Omega)} + \frac{|\alpha|}{\sqrt{2}\mu} \|\mathrm{curl}\, \boldsymbol{f}\|_{L^2(\Omega)} + \big(\frac{2\sqrt{2}}{\mu}\big)^{1+s} C_3 \|\boldsymbol{g}\|^{2+s}_{H^{1/2}(\partial\Omega)} \big)\,. \tag{4.27}$$

Note that these bounds still hold when α tends to zero.

These bounds, uniform with respect to h permit to establish weak convergence: we can extract a sub-sequence, still noted $\boldsymbol{u}_h, p_h, z_h$ such that

$$\lim_{h\to 0} \boldsymbol{u}_h = \boldsymbol{u} \ \text{ weakly in } W \ , \quad \lim_{h\to 0} p_h = p \ \text{ weakly in } L^2_0(\Omega) \ , \ \lim_{h\to 0} z_h = z \ \text{ weakly in } L^2(\Omega) \, .$$

Then we prove that $(\boldsymbol{u}, p)$ is the solution of (3.19)–(3.21) associated with z. But passing to the limit in (4.4), we do not recover (3.22), because in the trilinear terms of (4.4) we have the product of two weakly convergent sequences. For this, we need the strong convergence of the divergence. Nevertheless, by taking the difference between (4.2) and (3.19), we can prove that $\boldsymbol{u}_h$ converges strongly to $\boldsymbol{u}$ in $H^1(\Omega)^2$ and similarly p_h converges strongly to p in $L^2(\Omega)$. The strong convergence of $\boldsymbol{u}_h$ allows one to pass to the limit in (4.4) and we find (3.22).

It remains to establish error estimates. By taking the difference between the exact and discrete equations,we find the following equalities for all $\boldsymbol{v}_h$ in V_h, all q_h in M_h and all θ_h in Z_h:

$$\mu(\nabla(\boldsymbol{u}_h - \boldsymbol{u}), \nabla\, \boldsymbol{v}_h) + ((\boldsymbol{z}_h - \boldsymbol{z}) \times \boldsymbol{u}_h, \boldsymbol{v}_h) + (\boldsymbol{z} \times (\boldsymbol{u}_h - \boldsymbol{u}), \boldsymbol{v}_h) - (q_h - p, \operatorname{div} \boldsymbol{v}_h) = 0 \, , \tag{4.28}$$

$$\begin{aligned} \mu(z_h - z, \theta_h) + \alpha\, ((\boldsymbol{u}_h - \boldsymbol{u}) \cdot \nabla\, z_h, \theta_h) + \alpha\, (\boldsymbol{u} \cdot \nabla(z_h - z), \theta_h) + \frac{\alpha}{2}((\operatorname{div}(\boldsymbol{u}_h - \boldsymbol{u}))z_h, \theta_h) \\ = \mu(\operatorname{curl}(\boldsymbol{u}_h - \boldsymbol{u}), \theta_h) \, . \end{aligned} \tag{4.29}$$

From these, we deduce first estimates (note that $P_h(\boldsymbol{u}) - \boldsymbol{u}_h$ vanishes on $\partial\Omega$):

$$\begin{aligned} |\boldsymbol{u} - \boldsymbol{u}_h|_{H^1(\Omega)} \le 2|\boldsymbol{u} - P_h(\boldsymbol{u})|_{H^1(\Omega)} + \frac{S_4}{\mu}\|P_h(\boldsymbol{u})\|_{L^4(\Omega)}\|z - z_h\|_{L^2(\Omega)} + \frac{S_4}{\mu}\|z\|_{L^2(\Omega)}\|\boldsymbol{u} - P_h(\boldsymbol{u})\|_{L^4(\Omega)} \\ + \frac{\sqrt{2}}{\mu}\|p - r_h(p)\|_{L^2(\Omega)} \, , \end{aligned} \tag{4.30}$$

$$\begin{aligned} \|p - p_h\|_{L^2(\Omega)} \le 2\, \|p - r_h(p)\|_{L^2(\Omega)} + \frac{\mu}{\beta^*}|\boldsymbol{u} - P_h(\boldsymbol{u})|_{H^1(\Omega)} \\ + \frac{S_4\tilde{S}_4}{\beta^*}\left(\|z\|_{L^2(\Omega)}|\boldsymbol{u} - \boldsymbol{u}_h|_{H^1(\Omega)} + |\boldsymbol{u}_h|_{H^1(\Omega)}\|z - z_h\|_{L^2(\Omega)}\right) \, , \end{aligned} \tag{4.31}$$

where $r_h(p)$ is a good approximation of p in M_h (cf. [9], [20] or [56]);

$$\begin{aligned} \|z - z_h\|_{L^2(\Omega)} \le 2\, \|z - \lambda_h\|_{L^2(\Omega)} + \|\operatorname{curl}(\boldsymbol{u} - \boldsymbol{u}_h)\|_{L^2(\Omega)} \\ + \frac{|\alpha|}{\mu}\big(\|(\boldsymbol{u} - \boldsymbol{u}_h) \cdot \nabla\, \lambda_h\|_{L^2(\Omega)} + \|\boldsymbol{u} \cdot \nabla(z - \lambda_h)\|_{L^2(\Omega)} + \frac{1}{2}\|\operatorname{div}(\boldsymbol{u} - \boldsymbol{u}_h)\lambda_h\|_{L^2(\Omega)}\big) \, , \end{aligned} \tag{4.32}$$

for any $\lambda_h \in Z_h$. Clearly, the difficulty comes from the factors of $\frac{|\alpha|}{\mu}$ in (4.32). It does not seem possible to bound them without supposing that z belongs to $H^1(\Omega)$ and $\boldsymbol{u}$ to a smaller space than $H^1(\Omega)^2$ (compare with Remark 3.8). If we assume that $z \in H^1(\Omega)$ and $\boldsymbol{u} \in W^{1,2+1/4}(\Omega)^2$, we have:

$$\begin{aligned} \|z - z_h\|_{L^2(\Omega)} \le 2\, \|z - R_h(z)\|_{L^2(\Omega)} + \sqrt{2}|\boldsymbol{u} - \boldsymbol{u}_h|_{H^1(\Omega)} \\ + \frac{|\alpha|}{\mu}(|R_h(z)|_{H^1(\Omega)}(\|\boldsymbol{u} - \boldsymbol{u}_h\|_{L^\infty(\Omega)} + C\frac{\sqrt{2}}{2}|\boldsymbol{u} - \boldsymbol{u}_h|_{W^{1,2+1/4}(\Omega)}) + \|\boldsymbol{u}\|_{L^\infty(\Omega)}|z - R_h(z)|_{H^1(\Omega)}) \, , \end{aligned} \tag{4.33}$$

where $R_h(z)$ is also a good approximation of z in Z_h and C is a Sobolev imbedding constant. The choice of exponent $2 + 1/4$ is arbitrary; it suffices that this exponent be greater than two. But it is better to take it close to two, so as to avoid a quasi-uniformity condition on the mesh (i.e. $h_\kappa \ge \tau h$ for $\tau > 0$

independent of h). With this choice, we need a bound for $\boldsymbol{u}-\boldsymbol{u}_h$ in $W^{1,2+1/4}(\Omega)^2$; this is the object of (4.8):

$$|\boldsymbol{u}_h-\boldsymbol{u}|_{W^{1,r}(\Omega)} \le |P_h(\boldsymbol{u})-\boldsymbol{u}|_{W^{1,r}(\Omega)} + Ch^{2/r-1}(K_1(h)|P_h(\boldsymbol{u})-\boldsymbol{u}|_{H^1(\Omega)} + \|p-r_h(p)\|_{L^2(\Omega)}) + CK_2(h)\big(1+h^{2/r-1/2}(1+K_1(h))\big)\|z-z_h\|_{L^2(\Omega)}\,,$$

where

$$K_2(h)=\frac{1}{\mu}\|\boldsymbol{w}\|_{L^4(\Omega)}\big(1+\frac{S_4^2}{\mu}\|z\|_{L^2(\Omega)}\big)\,,$$

and $\boldsymbol{w}$ is the solution of the generalized Stokes equation (3.19)–(3.21) with z_h instead of z. It is written for a quasi-uniform triangulation, because it involves inverse inequalities; on the other hand, if $r\in[2,4]$, it requires no restriction on the angles of the domain. When the domain is convex (which is assumed in order that $z\in H^1(\Omega)$) and when $\boldsymbol{g}$ belongs to $H^{1/2+s}(\partial\Omega)^2$ for some $s\in(0,1/2)$, then this quasi-uniformity can be relaxed as follows: if

$$\forall\kappa\in\mathcal{T}_h\,,\ \rho_\kappa\ge\gamma\,h^6\,, \tag{4.34}$$

with a constant γ independente of h, then

$$\begin{aligned}|\boldsymbol{u}_h-\boldsymbol{u}|_{W^{1,2+1/4}(\Omega)} &\le |P_h(\boldsymbol{u})-\boldsymbol{u}|_{W^{1,2+1/4}(\Omega)} + \frac{C_1}{\rho_{\min}^{1/9}}(K_1(h)|P_h(\boldsymbol{u})-\boldsymbol{u}|_{H^1(\Omega)}\\ &+\frac{\sqrt{2}}{\mu}\|p-r_h(p)\|_{L^2(\Omega)}) + \|z-z_h\|_{L^2(\Omega)}\big(C_2K_2(h)+C_3h^{1/4}K_3(h)(1+K_1(h))\big)\,,\end{aligned} \tag{4.35}$$

where $\rho_{\min}$ is the minimum of ρ_κ for all κ in $\mathcal{T}_h$, C_i denote constants independent of h and

$$K_3(h)=\frac{1}{\mu}\left(\|\boldsymbol{w}\|_{L^\infty(\Omega)}+C_\infty K_2(h)\|z\|_{L^2(\Omega)}\right)\,.$$

The same estimate is valid for $\|\boldsymbol{u}_h-\boldsymbol{u}\|_{L^\infty(\Omega)}$ with the term $\|P_h(\boldsymbol{u})-\boldsymbol{u}\|_{L^\infty(\Omega)}$ in the right-hand side. By substituting these bounds into (4.33), supposing that the domain is convex and the data sufficiently small and smooth (a condition that is close to the one that guarantees uniqueness of the exact solution), we prove that $\|z-z_h\|_{L^2(\Omega)}$ satisfies an error inequality that shows that the scheme has *order one* when $z\in H^2(\Omega)$, $\boldsymbol{u}\in H^3(\Omega)^2$ and $p\in H^2(\Omega)$:

$$\|z-z_h\|_{L^2(\Omega)}+\|\boldsymbol{u}-\boldsymbol{u}_h\|_{H^1(\Omega)}+\|p-p_h\|_{L^2(\Omega)}\le C\,h\,.$$

4.2 An upwind approximation

The above result is not optimal with respect to the degree of the polynomials used. The loss of optimality arises from the discretization of the transport equation. We can gain a little – a factor $h^{1/2}$ – by an upwinding approximation of the transport term. There are several upwinding techniques. In [31] upwinding is achieved by "streamline diffusion", introduced by [40] (see also [41] and [52]). Whereas in [33] upwinding is produced by the discontinuous Galerkin method introduced by [43] for solving a neutron transport equation. We present here this second approach with the same velocity and pressure spaces as in the preceding paragraph. On the other hand, we take for Z_h piecewise polynomials of degree one in each triangle:

$$Z_h=\{\theta_h\in L^2(\Omega)\,;\,\forall\kappa\in\mathcal{T}_h\,,\ \theta_h|_\kappa\in I\!P_1\}\,. \tag{4.36}$$

For each discrete velocity $\boldsymbol{u}_h$ in $H^1_T(\Omega)$ and for each triangle κ, we set

$$\partial\kappa_-=\{\boldsymbol{x}\in\partial\kappa\,;\,\alpha\boldsymbol{u}_h\cdot\boldsymbol{n}<0\}\,. \tag{4.37}$$

Note that, when we describe all triangles κ of $\mathcal{T}_h$, $\partial\kappa_-$ only involves interior segments of $\mathcal{T}_h$ since $\boldsymbol{u}_h\cdot\boldsymbol{n}=0$ on $\partial\Omega$. Then, we discretize the trilinear terms $\alpha[(\boldsymbol{u}\cdot\nabla z,\theta)+\frac12((\operatorname{div}\boldsymbol{u})z,\theta)]$ by

$$c(\boldsymbol{u}_h;z_h,\theta_h)=\sum_{\kappa\in\mathcal{T}_h}\Big(\int_\kappa\alpha(\boldsymbol{u}_h\cdot\nabla z_h)\theta_h\,d\boldsymbol{x}+\int_{\partial\kappa_-}|\alpha\boldsymbol{u}_h\cdot\boldsymbol{n}|(z_h^{\rm int}-z_h^{\rm ext})\theta_h^{\rm int}ds\Big)+\frac{\alpha}{2}\int_\Omega(\operatorname{div}\boldsymbol{u}_h)z_h\theta_h\,d\boldsymbol{x}\,, \tag{4.38}$$

where the symbol int (resp. ext) denotes the trace on the segment $\partial\kappa$ of the function coming from the interior (resp. exterior) of κ. Note also that, when summing over all triangles, the integral runs over each internal side exactly once because $\boldsymbol{u}_h \cdot \boldsymbol{n}$ changes sign when passing from one triangle to the next adjacent triangle.

With Z_h defined by (4.36), the discrete scheme reads: *Find* $\boldsymbol{u}_h \in X_h + \boldsymbol{g}_h$, $p_h \in M_h$ *and* $z_h \in Z_h$ *solution of*

$$\forall \boldsymbol{v}_h \in X_h\,,\ \mu(\nabla\,\boldsymbol{u}_h, \nabla\,\boldsymbol{v}_h) + (\boldsymbol{z}_h \times \boldsymbol{u}_h, \boldsymbol{v}_h) - (p_h, \operatorname{div}\boldsymbol{v}_h) = (\boldsymbol{f}, \boldsymbol{v}_h)\,, \tag{4.39}$$

$$\forall q_h \in M_h\,,\ (q_h, \operatorname{div}\boldsymbol{u}_h) = 0\,, \tag{4.40}$$

$$\forall \theta_h \in Z_h\,,\ \mu\,(z_h, \theta_h) + c(\boldsymbol{u}_h; z_h, \theta_h) = \mu\,(\operatorname{curl}\boldsymbol{u}_h, \theta_h) + \alpha\,(\operatorname{curl}\boldsymbol{f}, \theta_h)\,. \tag{4.41}$$

Remark 4.3 We can also approximate z by piecewise constant functions in each triangle. It can be associated with the "mini-element" (cf. [2] or [29], [15]) for discretizing the velocity and pressure. The analysis below extends to this approximation and we find an error of the order of $h^{1/2}$. ∎

The numerical analysis of (4.39)–(4.41) is very close to that of (4.2)–(4.4), and we shall only present here modifications brought by (4.41). The following Green's formula is established in [43].

Lemma 4.4 *For all* $\boldsymbol{v}_h$ *in* X_h*, for all* z_h *and* θ_h *in* Z_h*, we have*

$$c(\boldsymbol{v}_h; z_h, \theta_h) = \sum_{\kappa\in\mathcal{T}_h} \Big(-\int_\kappa \alpha(\boldsymbol{v}_h \cdot \nabla\,\theta_h) z_h\,d\boldsymbol{x} + \int_{\partial\kappa_-} |\alpha\boldsymbol{v}_h \cdot \boldsymbol{n}| z_h^{\rm ext}(\theta_h^{\rm ext} - \theta_h^{\rm int}) ds\Big) - \frac{\alpha}{2}\int_\Omega (\operatorname{div}\boldsymbol{u}_h)\theta_h z_h\,d\boldsymbol{x}\,. \tag{4.42}$$

On one hand, when $\theta_h \in H^1(\Omega)$, (4.42) reduces to

$$c(\boldsymbol{v}_h; z_h, \theta_h) = -\int_\Omega \alpha(\boldsymbol{v}_h \cdot \nabla\,\theta_h) z_h\,d\boldsymbol{x} - \frac{\alpha}{2}\int_\Omega (\operatorname{div}\boldsymbol{u}_h)\theta_h z_h\,d\boldsymbol{x}\,. \tag{4.43}$$

On the other hand, when $\theta_h = z_h \in Z_h$, then

$$c(\boldsymbol{v}_h; z_h, z_h) = \frac{1}{2}\sum_{\kappa\in\mathcal{T}_h} \int_{\partial\kappa_-} |\alpha\boldsymbol{v}_h \cdot \boldsymbol{n}| (z_h^{\rm ext} - z_h^{\rm int})^2 ds\,. \tag{4.44}$$

This enables us to deduce from (4.41) the following bound:

$$\mu\,\|z_h\|^2_{L^2(\Omega)} + \frac{1}{2}\sum_{\kappa\in\mathcal{T}_h} \int_{\partial\kappa_-} |\alpha\boldsymbol{u}_h \cdot \boldsymbol{n}| (z_h^{\rm ext} - z_h^{\rm int})^2 ds = \mu\,(\operatorname{curl}\boldsymbol{u}_h, z_h) + \alpha\,(\operatorname{curl}\boldsymbol{f}, z_h)\,. \tag{4.45}$$

Whence the existence theorem:

Theorem 4.5 *There exists a constant* $C_1 > 0$*, independent of* h*, such that for all* $\mu > 0$ *and* $\alpha \in \mathbb{R}$*, for all* $\boldsymbol{f} \in H(curl, \Omega)$ *and all* $\boldsymbol{g} \in H^{1/2}(\partial\Omega)^2$ *satisfying* $\boldsymbol{g} \cdot \boldsymbol{n} = 0$*, if*

$$h_b < C_1 \mu^{2+\tau} \|\boldsymbol{g}\|^{-2-\tau}_{H^{1/2}(\partial\Omega)}\,,\ \ \textit{for some } \tau > 0\,, \tag{4.46}$$

then (4.39)–(4.41) has at least one solution and each solution satisfies the a priori estimates (4.22), (4.23),

$$\|z_h\|_{L^2(\Omega)} \le \sqrt{2}|\boldsymbol{u}_h|_{H^1(\Omega)} + \frac{|\alpha|}{\mu}\|\operatorname{curl}\boldsymbol{f}\|_{L^2(\Omega)}\,, \tag{4.47}$$

$$\frac{1}{2}\sum_{\kappa\in\mathcal{T}_h} \int_{\partial\kappa_-} |\alpha\boldsymbol{u}_h \cdot \boldsymbol{n}| \big(z_h^{\rm ext} - z_h^{\rm int}\big)^2 ds \le \big(\sqrt{2}\mu|\boldsymbol{u}_h|_{H^1(\Omega)} + |\alpha|\|\operatorname{curl}\boldsymbol{f}\|_{L^2(\Omega)}\big)\|z_h\|_{L^2(\Omega)}\,. \tag{4.48}$$

Moreover, we have for all real number $s > \frac{\tau}{2}$*:*

$$\|z_h\|_{L^2(\Omega)} \le \frac{C_2}{\mu^{1+s}}\|\boldsymbol{g}\|^{2+s}_{H^{1/2}(\partial\Omega)} + 2\sqrt{2}\frac{S_2}{\mu}\|\boldsymbol{f}\|_{L^2(\Omega)} + 2\frac{|\alpha|}{\mu}\|\operatorname{curl}\boldsymbol{f}\|_{L^2(\Omega)}\,, \tag{4.49}$$

where C_2 *depends on* s *and* τ*, but not on* h *or* μ*.*

As in the preceding paragraph, we prove that a subsequence of $(\boldsymbol{u}_h, p_h, z_h)$ converges weakly to functions $(\boldsymbol{u}, p, z)$ in $H^1_T(\Omega) \times L^2_0(\Omega) \times L^2(\Omega)$; in addition, $\sum_{\kappa\in\mathcal{T}_h} \int_{\partial\kappa_-} |\alpha \boldsymbol{u}_h \cdot \boldsymbol{n}|(z_h^{\rm ext} - z_h^{\rm int})^2 ds$ converges to a number $S \geq 0$. Passing to the limit in (4.39), we see that $(\boldsymbol{u}, p, z)$ satisfies (3.19). Next, we prove strong convergence of $\boldsymbol{u}_h$ and p_h, and the strong convergence of $\boldsymbol{u}_h$ allows one to pass to the limit in (3.22). Next, we show that

$$\lim_{h\to 0} \|z_h\|_{L^2(\Omega)} = \|z\|_{L^2(\Omega)} .$$

This implies on one hand the strong convergence of z_h and on the other hand that $S = 0$.

As far as error estimates are concerned, (4.30) and (4.31) are unchanged, while(4.32) is replaced by:

$$\begin{aligned}
&\mu\|z_h - \zeta_h\|^2_{L^2(\Omega)} + \sum_{\kappa\in\mathcal{T}_h} \frac{1}{2} \int_{\partial\kappa_-} |\alpha \boldsymbol{u}_h \cdot \boldsymbol{n}|((z_h - \zeta_h)^{\rm ext} - (z_h - \zeta_h)^{\rm int})^2 ds \\
&\quad + \sum_{\kappa\in\mathcal{T}_h} \left(-\alpha \int_\kappa \boldsymbol{u}_h \cdot \nabla(z_h - \zeta_h)(\zeta_h - z) d\boldsymbol{x} + \int_{\partial\kappa_-} |\alpha \boldsymbol{u}_h \cdot \boldsymbol{n}|((z_h - \zeta_h)^{\rm ext} - (z_h - \zeta_h)^{\rm int})(\zeta_h - z)^{\rm ext} ds \right) \\
&\quad - \frac{\alpha}{2} \int_\Omega \mathrm{div}(\boldsymbol{u}_h - \boldsymbol{u})(\zeta_h - z)(z_h - \zeta_h) d\boldsymbol{x} + \frac{\alpha}{2} \int_\Omega \mathrm{div}(\boldsymbol{u}_h - \boldsymbol{u}) z (z_h - \zeta_h) d\boldsymbol{x} \\
&\quad + \alpha \int_\Omega (\boldsymbol{u}_h - \boldsymbol{u}) \cdot \nabla z (z_h - \zeta_h) d\boldsymbol{x} = \mu(z - \zeta_h, z_h - \zeta_h) + \mu(\mathrm{curl}(\boldsymbol{u}_h - \boldsymbol{u}), z_h - \zeta_h) ,
\end{aligned} \tag{4.50}$$

for arbitrary $\zeta_h \in Z_h$. The third and fourth terms of the left-hand side must be handled with care in order to take advantage of the upwinding (cf. [52]). Here we take advantage of the discontinuity of functions of Z_h and we obtain with $\varrho_h(z)$, the L^2 projection of z in $I\!P_1$ in each triangle κ:

$$\begin{aligned}
\|z_h - \varrho_h(z)\|^2_{L^2(\Omega)} &\leq c_5 \frac{|\alpha|^2}{\mu^2} \big(|z|^2_{H^1(\Omega)} \|\boldsymbol{u}_h - \boldsymbol{u}\|^2_{L^\infty(\Omega)} + |\boldsymbol{u}|^2_{W^{1,\infty}(\Omega)} \|z - \varrho_h(z)\|^2_{L^2(\Omega)} \big) \\
&\quad + c_6 h \frac{|\alpha|}{\mu} \|\boldsymbol{u}_h\|_{L^\infty(\Omega)} |z - \varrho_h(z)|^2_{H^1(\Omega)} + c_7 \frac{|\alpha|^2}{\mu^2} \|z\|^2_{H^1(\Omega)} \|\mathrm{div}(\boldsymbol{u}_h - \boldsymbol{u})\|^2_{L^{2+1/4}(\Omega)} \\
&\quad + 2\big(\|z - \varrho_h(z)\|^2_{L^2(\Omega)} + 2|\boldsymbol{u}_h - \boldsymbol{u}|^2_{H^1(\Omega)} \big) .
\end{aligned} \tag{4.51}$$

Besides this, (4.35) still holds here under the same hypotheses. Hence, as in the preceding paragraph, we arrive at the following error inequality, under similar assumptions on the domain, the data and the triangulation:

$$\begin{aligned}
\|z_h - z\|_{L^2(\Omega)} \leq C\big(&\|P_h(\boldsymbol{u}) - \boldsymbol{u}\|_{L^\infty(\Omega)} + |P_h(\boldsymbol{u}) - \boldsymbol{u}|_{W^{1,2+1/4}(\Omega)} + \frac{1}{\rho_{\min}^{1/9}} (|P_h(\boldsymbol{u}) - \boldsymbol{u}|_{H^1(\Omega)} \\
&+ \|r_h(p) - p\|_{L^2(\Omega)}) + \|P_h(\boldsymbol{u}) - \boldsymbol{u}\|_{L^4(\Omega)} + \|\varrho_h(z) - z\|_{L^2(\Omega)} + h^{1/2} |\varrho_h(z) - z|_{H^1(\Omega)} \big) .
\end{aligned} \tag{4.52}$$

The last term above is dominating. In the best case, it is of the order of $h^{3/2}$ and improving it appears problematic. As a conlusion, if $z \in H^2(\Omega)$, $\boldsymbol{u} \in H^3(\Omega)^2$ and $p \in H^2(\Omega)$, the scheme has order 3/2:

$$|\boldsymbol{u}_h - \boldsymbol{u}|_{H^1(\Omega)} + \|p_h - p\|_{L^2(\Omega)} + \|z_h - z\|_{L^2(\Omega)} \leq C\, h^{3/2} . \tag{4.53}$$

We end this paragraph with a successive approximation algorithm for computing numerically $(\boldsymbol{u}_h, p_h, z_h)$ solution of (4.39)–(4.41). Let z_h^0 be an arbitrary function of Z_h (for instance $z_h^0 = 0$); for $k \geq 0$ we compute the sequence $\boldsymbol{u}_h^k \in X_h + \boldsymbol{g}_h$, $p_h^k \in M_h$ and $z_h^{k+1} \in Z_h$ by solving first the generalized Stokes problem:

$$\forall \boldsymbol{v}_h \in X_h , \ \mu(\nabla \boldsymbol{u}_h^k, \nabla \boldsymbol{v}_h) + (\boldsymbol{z}_h^k \times \boldsymbol{u}_h^k, \boldsymbol{v}_h) - (p_h^k, \mathrm{div}\, \boldsymbol{v}_h) = (\boldsymbol{f}, \boldsymbol{v}_h) ,$$

$$\forall q_h \in M_h , \ (q_h, \mathrm{div}\, \boldsymbol{u}_h^k) = 0 ,$$

and next the linear transport equation:

$$\forall \theta_h \in Z_h\,,\ \mu\,(z_h^{k+1},\theta_h) + c(\boldsymbol{u}_h^k; z_h^{k+1},\theta_h) = \mu\,(\operatorname{curl}\boldsymbol{u}_h^k,\theta_h) + \alpha\,(\operatorname{curl}\boldsymbol{f},\theta_h)\,.$$

We prove that this sequence of functions satisfy a priori estimates similar to (4.22), (4.23), (4.48) and (4.49), uniform with respect to h and k. Therefore, we can extract a subsequence that converges uniformly with respect to h when k tends to infinity. Under the sufficient conditions of Theorem 3.4 [33] on the domain and data, the limiting functions is a solution $(\boldsymbol{u}_h, p_h, z_h)$ of (4.39)–(4.41). We can proceed in the same fashion to compute numerically a solution of problem (4.2)–(4.4).

Remark 4.6 Compare the convergence of this algorithm with the results of [4] obtained for the same problem, with the same finite-element spaces, but with the other formulation (cf. (3.16)–(3.18)). The authors prove existence and convergence of a solution of their scheme, but cannot establish convergence of their successive approximation algorithm, except if the algorithm starts with an approximation of the exact solution that has the same order of accuracy as their final result. In other words, they must have already solved the problem they propose to solve in order to guarantee convergence of their algorithm. This is a popular approach; it can be found in all publications intending to perform the numerical analysis of Oldroyd models. All establish existence and convergence of a discrete solution, but none of them knows how to prove convergence of an algorithm to compute this solution. The reader can refer to the article by Picasso and Rappaz [51] where the analysis is an application of the Implicit Function Theorem. ∎

4.3 Heuristic remarks on approximation in three dimensions

To simplify, consider the problem with a zero Dirichlet boundary condition in 3-D: *Find* $(\boldsymbol{u}, p, \boldsymbol{z})$ *in* $W^{1,\infty}(\Omega)^3 \times L^2_0(\Omega) \times L^2(\Omega)^3$ *solution of*

$$-\mu\,\Delta\,\boldsymbol{u} + \boldsymbol{z}\times\boldsymbol{u} + \nabla\,p = \boldsymbol{f}\ ,\ \operatorname{div}\boldsymbol{u} = 0\,,$$

$$\boldsymbol{u} = \boldsymbol{0} \quad \text{on } \partial\Omega\,,$$

$$\mu\,\boldsymbol{z} + \alpha\,\boldsymbol{u}\cdot\nabla\,\boldsymbol{z} - \alpha\,\boldsymbol{z}\cdot\nabla\,\boldsymbol{u} = \mu\,\mathbf{curl}\,\boldsymbol{u} + \alpha\,\mathbf{curl}\,\boldsymbol{f}\,.$$

It is equivalent to the steady-state version of (3.1)–(3.4). The theoretical analysis of [7] shows that, under suitable restrictions on the data and the domain, this problem has at least one solution.

Consider the centered approximation in 3-D with the Taylor-Hood $I\!P_2$-$I\!P_1$ finite-element spaces:

$$X_h = \{\boldsymbol{v}_h \in \mathcal{C}^0(\overline{\Omega})^3\,;\,\forall\kappa\in\mathcal{T}_h\,,\,\boldsymbol{v}_h|_\kappa \in I\!P_2^3\,,\,\boldsymbol{v}_h|_{\partial\Omega} = \boldsymbol{0}\}\,,$$

$$M_h = \{q_h \in \mathcal{C}^0(\overline{\Omega})\,;\,\forall\kappa\in\mathcal{T}_h\,,\,q_h|_\kappa \in I\!P_1\} \cap L^2_0(\Omega)\,,$$

$$Z_h = \{\boldsymbol{\theta}_h \in \mathcal{C}^0(\overline{\Omega})^3\,;\,\forall\kappa\in\mathcal{T}_h\,,\,\boldsymbol{\theta}_h|_\kappa \in I\!P_1^3\}\,.$$

The pair (X_h, M_h) satisfies a uniform inf-sup condition [15], but in contrast to the 2-D element, it does not appear to have a quasi-local operator unless the mesh is appropriately structured. If the mesh is made of hexahedra decomposed into twelve tetrahedra, then an operator P_h can be constructed as in two dimensions and it has the same approximation properties. With these spaces, we discretize *Problem P* by: *Find* $\boldsymbol{u}_h$ *in* X_h, p_h *in* M_h *and* $\boldsymbol{z}_h$ *in* Z_h, *such that*

$$\forall \boldsymbol{v}_h \in X_h\,,\ \mu(\nabla\,\boldsymbol{u}_h, \nabla\,\boldsymbol{v}_h) + (\boldsymbol{z}_h\times\boldsymbol{u}_h, \boldsymbol{v}_h) - (p_h, \operatorname{div}\boldsymbol{v}_h) = (\boldsymbol{f}, \boldsymbol{v}_h)\,, \tag{4.54}$$

$$\forall q_h \in M_h\,,\ (q_h, \operatorname{div}\boldsymbol{u}_h) = 0\,, \tag{4.55}$$

$$\begin{aligned}\forall \boldsymbol{\theta}_h \in Z_h\,,\ \mu\,(\boldsymbol{z}_h, \boldsymbol{\theta}_h) + \alpha[(\boldsymbol{u}_h\cdot\nabla\,\boldsymbol{z}_h - \boldsymbol{z}_h\cdot\nabla\,\boldsymbol{u}_h, \boldsymbol{\theta}_h) + \frac{1}{2}((\operatorname{div}\boldsymbol{u}_h)\boldsymbol{z}_h, \boldsymbol{\theta}_h)] \\ = \mu\,(\mathbf{curl}\,\boldsymbol{u}_h, \boldsymbol{\theta}_h) + \alpha\,(\mathbf{curl}\,\boldsymbol{f}, \boldsymbol{\theta}_h)\,.\end{aligned} \tag{4.56}$$

Heuristically speaking, to obtain a priori estimates for $\boldsymbol{u}_h, p_h, \boldsymbol{z}_h$, we must first derive from (4.54) with fixed $\boldsymbol{z}_h$, a bound of the form:

$$\|\nabla\,\boldsymbol{u}_h\|_{L^\infty(\Omega)} \le C\,,$$

where C is uniformly bounded with respect to h. In 2-D, this inequality stems solely from inverse inequalites because H^2 is "almost imbedded" into $W^{1,\infty}$. But in 3-D, this imbedding fails by a lot, and we require sharp approximation properties of the Stokes projection. First, reverting to the solution $\boldsymbol{w}$ of the generalized Stokes equation (3.19)–(3.21) with $\boldsymbol{z}_h$ instead of $\boldsymbol{z}$, we can prove that if the domain is convex and the triangulation quasi-uniform, in view of the good approximation properties of P_h:

$$\|\boldsymbol{u}_h - P_h(\boldsymbol{w})\|_{L^\infty(\Omega)} \leq C\|\boldsymbol{z}_h\|_{L^2(\Omega)}\,,$$

where C is a constant independent of h. Next, for fixed $\boldsymbol{u}_h$ and $\boldsymbol{z}_h$, we see that (4.54) is a discretization of:

$$-\Delta\boldsymbol{v} + \nabla\, q = \boldsymbol{f} - \boldsymbol{z}_h \times \boldsymbol{u}_h\ ,\ \operatorname{div}\boldsymbol{v} = 0\ \ \text{in}\ \Omega\,,$$
$$\boldsymbol{v} = \boldsymbol{0}\ \ \text{on}\ \partial\Omega\,;$$

more precisely, $\boldsymbol{u}_h$ is the Stokes projection of $\boldsymbol{v}$. Owing to Sobolev's imbeddings, we see that if the angles of the domain are sufficiently restricted for the homogeneous Stokes problem to have the regularity: $L^r(\Omega)^3$ gives $W^{2,r}(\Omega)^3$ for some $r > 3$, and if $\boldsymbol{f} \in L^r(\Omega)^3$, then $\boldsymbol{v}$ and q satisfy

$$|\boldsymbol{v}|_{W^{2,r}(\Omega)} + |q|_{W^{1,r}(\Omega)} \leq C\|\boldsymbol{z}_h\|_{L^2(\Omega)}\|\boldsymbol{z}_h\|_{L^r(\Omega)}\,.$$

Therefore, the estimates we need reduce on one hand to proving

$$|\boldsymbol{u}_h|_{W^{1,\infty}(\Omega)} \leq C(|\boldsymbol{v}|_{W^{2,r}(\Omega)} + |q|_{W^{1,r}(\Omega)})\,, \tag{4.57}$$

with a constant C independent of h and on the other hand, to proving that for given $\boldsymbol{u}_h$ in $W^{1,\infty}(\Omega)^3$ and $\mathbf{curl}\,\boldsymbol{f}$ in $L^r(\Omega)^3$, the solution $\boldsymbol{z}_h$ of (4.56) satisfies:

$$\|\boldsymbol{z}_h\|_{L^r(\Omega)} \leq C\,, \tag{4.58}$$

for the above $r > 3$. The proof of (4.57) is very recent while (4.58) is *an open problem.*

Remark 4.7 Regarding the exact problem, if $\boldsymbol{f} \in L^r(\Omega)^3$ and the domain is as above, the solution $\boldsymbol{u}$ of the first equation, for $\boldsymbol{z}$ given in $L^r(\Omega)^3$, satisfies the estimate:

$$|\boldsymbol{u}|_{W^{2,r}(\Omega)} \leq C_1\big(\|\boldsymbol{f}\|_{L^r(\Omega)} + C_2\|\boldsymbol{z}\|_{L^r(\Omega)} + C_3\|\boldsymbol{z}\|^2_{L^r(\Omega)}\big)\,.$$

On the other hand, if $\mathbf{curl}\,\boldsymbol{f} \in L^r(\Omega)^3$, the solution $\boldsymbol{z}$ of the second equation, for $\boldsymbol{u}$ given in $W^{1,\infty}(\Omega)^3$ verifying (3.37):

$$|\alpha|\|\nabla\,\boldsymbol{u}\|_{L^\infty(\Omega)} \leq \mu - \eta\ \ \text{for some real number}\ \eta > 0, \eta < \mu\,,$$

is bounded by:

$$\|\boldsymbol{z}\|_{L^r(\Omega)} \leq \frac{1}{\eta}\big(\mu\|\mathbf{curl}\,\boldsymbol{u}\|_{L^r(\Omega)} + |\alpha|\|\mathbf{curl}\,\boldsymbol{f}\|_{L^r(\Omega)}\big)\,.$$

Therefore, we can reasonably hope to be able to prove (4.58). ∎

5 A least-squares algorithm

The algorithm described here is published in the thesis of Park [49]. It applies to the steady and unsteady problem in two and three dimensions, with a homogeneous Dirichlet boundary condition; it could also apply to a tangential condition (3.21). To simplify, we restrict the discussion to the steady problem.

Revert to the steady version of (3.1)–(3.4):

$$-\mu\,\Delta\,\boldsymbol{u} + \mathbf{curl}(\boldsymbol{u} - \alpha\,\Delta\,\boldsymbol{u}) \times \boldsymbol{u} + \nabla\, p = \boldsymbol{f}\ ,\ \operatorname{div}\boldsymbol{u} = 0 \quad \text{in}\ \Omega\,, \tag{5.1}$$

$$\boldsymbol{u} = \boldsymbol{0} \quad \text{on}\ \partial\Omega\,. \tag{5.2}$$

In three dimensions, we suppose that the data are such that this problem has at least one solution.

Instead of $\boldsymbol{z}$, we introduce the auxiliary variable $\boldsymbol{w}$ formally defined by:

$$\boldsymbol{w} = \boldsymbol{u} - \alpha\,\Delta\,\boldsymbol{u}\,.$$

Substituting into (5.1), we obtain:

$$-\mu\,\Delta\,\boldsymbol{u} + \mathbf{curl}\,\boldsymbol{w} \times \boldsymbol{u} + \nabla\,p = \boldsymbol{f}\,.$$

Note that $\boldsymbol{w}$ determines $\boldsymbol{u} \in V$ in each of the last two equations. Hence, we cannot use both for an iterative method; but in contrast, we can use them for a least-squares method. Whence the following algorithm:

- *For* $\boldsymbol{w}$ *given in* $H^1(\Omega)^d$, *find* $\boldsymbol{u}_1 = \boldsymbol{u}_1(\boldsymbol{w})$ *in* $H_0^1(\Omega)^d$ *solution of*

$$\boldsymbol{u}_1 - \alpha\,\Delta\,\boldsymbol{u}_1 = \boldsymbol{w} \quad \text{in } \Omega\,. \tag{5.3}$$

- *For the same* $\boldsymbol{w}$, *find* $(\boldsymbol{u}_2 = \boldsymbol{u}_2(\boldsymbol{w}), p)$ *in* $H_0^1(\Omega)^d \times L_0^2(\Omega)$ *solution of*

$$-\mu\,\Delta\,\boldsymbol{u}_2 + \mathbf{curl}\,\boldsymbol{w} \times \boldsymbol{u}_2 + \nabla\,p = \boldsymbol{f}\ ,\ \mathrm{div}\,\boldsymbol{u}_2 = 0 \quad \text{in } \Omega\,, \tag{5.4}$$

The systems (5.3) and (5.4) can be put into variational formulations and each one has a unique solution, but clearly there is no reason why $\boldsymbol{u}_1(\boldsymbol{w}) = \boldsymbol{u}_2(\boldsymbol{w})$. This equality is "forced" by finding $\boldsymbol{w} \in H^1(\Omega)^d$ that minimizes the norm $|\boldsymbol{u}_1(\boldsymbol{w}) - \boldsymbol{u}_2(\boldsymbol{w})|_{H^1(\Omega)}$. Thus, we introduce the functional $J : H^1(\Omega)^d \mapsto I\!R$ defined by:

$$\forall \boldsymbol{v} \in H^1(\Omega)^d\,,\ J(\boldsymbol{v}) = \frac{1}{2}|\boldsymbol{u}_1(\boldsymbol{v}) - \boldsymbol{u}_2(\boldsymbol{v})|^2_{H^1(\Omega)}\,, \tag{5.5}$$

and we rewrite the problem as: *Find* $\boldsymbol{w}^* \in H^1(\Omega)^d$ *such that*

$$J(\boldsymbol{w}^*) = \inf_{\boldsymbol{v} \in H^1(\Omega)^d} J(\boldsymbol{v})\,. \tag{5.6}$$

It follows from the above assumptions that the original problem (5.1)–(5.2) has at least one solution; hence the minimum of $J(\boldsymbol{w})$ is attained and is zero.

We propose to approximate $\boldsymbol{w}^*$ by a simple gradient algorithm:

- Starting step: guess $\boldsymbol{w}^0 \in H^1(\Omega)^d$ and choose a threshold ε.
- General step: for $n \geq 0$, knowing $\boldsymbol{w}^n$ and while

$$J(\boldsymbol{w}^n) > \varepsilon\,, \tag{5.7}$$

compute the gradient $\boldsymbol{g}^n \in H^1(\Omega)^d$ solution of

$$\forall \boldsymbol{h} \in H^1(\Omega)^d\,,\ (\boldsymbol{g}^n, \boldsymbol{h}) + (\nabla\,\boldsymbol{g}^n, \nabla\,\boldsymbol{h}) = J'(\boldsymbol{w}^n) \cdot \boldsymbol{h}\,, \tag{5.8}$$

and the number ρ_n by

$$J(\boldsymbol{w}^n - \rho_n \boldsymbol{g}^n) = \inf_{\rho \in I\!R} J(\boldsymbol{w}^n - \rho \boldsymbol{g}^n)\,, \tag{5.9}$$

then compute $\boldsymbol{w}^{n+1}$ by

$$\boldsymbol{w}^{n+1} = \boldsymbol{w}^n - \rho_n \boldsymbol{g}^n\,, \tag{5.10}$$

and return to the general step.

Let us describe the computations. To avoid using the curl operator explicitly in (5.4), we use:

$$\forall \boldsymbol{u}, \boldsymbol{v} \in H_0^1(\Omega)^d\,,\ (\mathbf{curl}\,\boldsymbol{w} \times \boldsymbol{u}, \boldsymbol{v}) = b(\boldsymbol{v}; \boldsymbol{u}, \boldsymbol{w}) - b(\boldsymbol{u}; \boldsymbol{v}, \boldsymbol{w}) - (\nabla(\boldsymbol{u} \cdot \boldsymbol{w}), \boldsymbol{v})\,,$$

where

$$b(\boldsymbol{u};\boldsymbol{v},\boldsymbol{w}) = \sum_{i,j=1}^{d} u_i \frac{\partial v_j}{\partial x_i} w_j \, d\boldsymbol{x} \,.$$

Then (5.3) and (5.4) are respectively equivalent to: Find $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ in $H_0^1(\Omega)^d$ solutions of

$$\forall \boldsymbol{v} \in H_0^1(\Omega)^d \,,\, (\boldsymbol{u}_1, \boldsymbol{v}) + \alpha(\nabla \boldsymbol{u}_1, \nabla \boldsymbol{v}) = (\boldsymbol{w}, \boldsymbol{v}) \,, \tag{5.11}$$

$$\forall \boldsymbol{v} \in H_0^1(\Omega)^d \,,\, \mu \,(\nabla \boldsymbol{u}_2, \nabla \boldsymbol{v}) - b(\boldsymbol{u}_2; \boldsymbol{v}, \boldsymbol{w}) + b(\boldsymbol{v}; \boldsymbol{u}_2, \boldsymbol{w}) - (\tilde{p}, \boldsymbol{v}) = (\boldsymbol{f}, \boldsymbol{v}) \,, \tag{5.12}$$

$$\forall q \in L_0^2(\Omega) \,,\, (q, \operatorname{div} \boldsymbol{u}_2) = 0 \,, \tag{5.13}$$

where up to a constant, $\tilde{p} = p - \boldsymbol{u}_2 \cdot \boldsymbol{w}$. The first is a system of d decoupled Laplace equations and $\boldsymbol{u}_1$ depends linearly on $\boldsymbol{w}$. The second is a linearized Navier-Stokes problem, but the dependence of $\boldsymbol{u}_2$ on $\boldsymbol{w}$ is not quadratic. Set

$$\boldsymbol{u}_1 = D_\alpha^{-1}(\boldsymbol{w}) \,,$$

and $H(\boldsymbol{w}) = \boldsymbol{u}_1(\boldsymbol{w}) - \boldsymbol{u}_2(\boldsymbol{w})$; we can write:

$$H'(\boldsymbol{w}) \cdot \boldsymbol{h} = \boldsymbol{U}_1 - \boldsymbol{U}_2 \,,$$

where

$$\boldsymbol{U}_1 = D_\alpha^{-1}(\boldsymbol{h}) \,,$$

and $\boldsymbol{U}_2 \in V$ is the solution of

$$\forall \boldsymbol{v} \in V \,,\, \mu \,(\nabla \boldsymbol{U}_2, \nabla \boldsymbol{v}) - b(\boldsymbol{U}_2; \boldsymbol{v}, \boldsymbol{w}) + b(\boldsymbol{v}; \boldsymbol{U}_2, \boldsymbol{w}) = b(\boldsymbol{u}_2; \boldsymbol{v}, \boldsymbol{h}) - b(\boldsymbol{v}; \boldsymbol{u}_2, \boldsymbol{h}) \,. \tag{5.14}$$

It is readily seen that problem (5.14) has a unique solution, because the bilinear form in the left-hand side is trivially elliptic and is continuous for $\boldsymbol{w}$ in $L^4(\Omega)^d$. The analysis done in Section 2 yields on one hand,

$$|\boldsymbol{u}_2|_{H^1(\Omega)} \leq \frac{S_2}{\mu} \|\boldsymbol{f}\|_{L^2(\Omega)} \,, \tag{5.15}$$

and on the other hand,

$$|\boldsymbol{u}_2(\boldsymbol{h}_1) - \boldsymbol{u}_2(\boldsymbol{h}_2)|_{H^1(\Omega)} \leq 2 \frac{S_2 S_4}{\mu^2} \|\boldsymbol{f}\|_{L^2(\Omega)} \|\boldsymbol{h}_1 - \boldsymbol{h}_2\|_{L^4(\Omega)} \,. \tag{5.16}$$

Similarly,

$$|\boldsymbol{u}_2'(\boldsymbol{w}) \cdot \boldsymbol{h}|_{H^1(\Omega)} \leq 2 \frac{S_2 S_4}{\mu^2} \|\boldsymbol{f}\|_{L^2(\Omega)} \|\boldsymbol{h}\|_{L^4(\Omega)} \,. \tag{5.17}$$

None of the higher-order derivatives of $\boldsymbol{u}_2$ vanish; their behaviour is analogous to that of the first derivative. Thus, the third derivative does not vanish, so that $\boldsymbol{u}_2$ is not quadratic with respect to $\boldsymbol{w}$.

Now we turn to the gradient $\boldsymbol{g}$. It is defined by: $\boldsymbol{g} \in H^1(\Omega)^d$, solution of

$$\forall \boldsymbol{v} \in H^1(\Omega)^d \,,\, (\boldsymbol{g}, \boldsymbol{v}) + (\nabla \boldsymbol{g}, \nabla \boldsymbol{v}) = (\nabla(\boldsymbol{U}_1 - \boldsymbol{U}_2), \nabla(\boldsymbol{u}_1 - \boldsymbol{u}_2)) \,.$$

But, as the divergence of $\boldsymbol{u}_1$ is not zero, we define the projection $P\boldsymbol{u}_1 \in V$ of $\boldsymbol{u}_1$ on V:

$$\forall \boldsymbol{v} \in V \,,\, (\nabla P\boldsymbol{u}_1, \nabla \boldsymbol{v}) = (\nabla \boldsymbol{u}_1, \nabla \boldsymbol{v})) \,,$$

and the gradient can also be expressed as

$$\forall \boldsymbol{v} \in H^1(\Omega)^d \,,\, (\boldsymbol{g}, \boldsymbol{v}) + (\nabla \boldsymbol{g}, \nabla \boldsymbol{v}) = (\nabla \boldsymbol{U}_1, \nabla(\boldsymbol{u}_1 - \boldsymbol{u}_2)) - (\nabla \boldsymbol{U}_2, \nabla(P\boldsymbol{u}_1 - \boldsymbol{u}_2)) \,. \tag{5.18}$$

With (5.14), the equation of the gradient becomes:

$$\forall \boldsymbol{h} \in H^1(\Omega)^d\,,\ (\boldsymbol{g},\boldsymbol{h}) + (\nabla\,\boldsymbol{g},\nabla\,\boldsymbol{h}) = \frac{1}{\alpha}(\boldsymbol{h},\boldsymbol{u}_1-\boldsymbol{u}_2) - \frac{1}{\alpha}(\boldsymbol{U}_1,\boldsymbol{u}_1-\boldsymbol{u}_2) \\ - \frac{1}{\mu}\big[b(\boldsymbol{U}_2;P\boldsymbol{u}_1-\boldsymbol{u}_2,\boldsymbol{w}) + b(\boldsymbol{u}_2;P\boldsymbol{u}_1,\boldsymbol{w}) - b(P\boldsymbol{u}_1;\boldsymbol{u}_2,\boldsymbol{h}) - b(P\boldsymbol{u}_1-\boldsymbol{u}_2;\boldsymbol{U}_2,\boldsymbol{w})\big]\,. \tag{5.19}$$

Now, we must compute ρ according to (5.9). This computation is heuristic, because $\boldsymbol{u}_2'''$ is not zero. The equation

$$(\nabla\,H'(\boldsymbol{w}^m-\rho\boldsymbol{g}^m)\cdot\boldsymbol{g}^m, \nabla\,H(\boldsymbol{w}^m-\rho\boldsymbol{g}^m)) = 0\,, \tag{5.20}$$

determines ρ. Let us approximate $H(\boldsymbol{w}^m-\rho\boldsymbol{g}^m)$ by its second-order expansion:

$$H(\boldsymbol{w}^m-\rho\boldsymbol{g}^m) \simeq H(\boldsymbol{w}^m) - \rho H'(\boldsymbol{w}^m)\cdot\boldsymbol{g}^m + \frac{\rho^2}{2}H''(\boldsymbol{w}^m)\cdot(\boldsymbol{g}^m,\boldsymbol{g}^m)\,,$$

and similarly

$$H'(\boldsymbol{w}^m-\rho\boldsymbol{g}^m)\cdot\boldsymbol{g}^m \simeq H(\boldsymbol{w}^m)\cdot\boldsymbol{g}^m - \rho H''(\boldsymbol{w}^m)\cdot(\boldsymbol{g}^m,\boldsymbol{g}^m)\,.$$

By substituting these two approximations into (5.20), we find an equation of the form $\varphi(\rho)=0$ where φ is a polynomial of degree three:

$$\varphi(\rho) = \|\boldsymbol{g}^m\|^2_{H^1(\Omega)} - \rho\big[(\nabla\,H''(\boldsymbol{w}^m)\cdot(\boldsymbol{g}^m,\boldsymbol{g}^m),\nabla\,H(\boldsymbol{w}^m)) + |H'(\boldsymbol{w}^m)\cdot\boldsymbol{g}^m|^2_{H^1(\Omega)}\big] \\ \frac{3}{2}\rho^2(\nabla\,H'(\boldsymbol{w}^m)\cdot\boldsymbol{g}^m,\nabla\,H''(\boldsymbol{w}^m)\cdot(\boldsymbol{g}^m,\boldsymbol{g}^m)) - \frac{1}{2}\rho^3|H''(\boldsymbol{w}^m)\cdot(\boldsymbol{g}^m,\boldsymbol{g}^m)|^2_{H^1(\Omega)}\,.$$

A study of φ enables us to locate and approximate its roots.

For the moment, we do not know how to establish convergence of this algorithm, even under strong hypotheses on the data, such as the sufficient conditions for uniqueness. The difficulty lies in the non-quadratic dependence of $\boldsymbol{u}_2$ on $\boldsymbol{w}$. Nevertheless the numerical experiments of [49] give good results. The velocity and pressure are approximated by the classical Taylor-Hood finite-element spaces that we have seen before: X_h is defined by (4.9) and M_h by (4.10). The variable $\boldsymbol{w}$ is approximated in the space Z_h defined by (4.11).

References

[1] Adams, R. A., *Sobolev Spaces*, Academic Press, New York, NY, 1975.

[2] Arnold, D., Brezzi, F. and Fortin, M., *A stable finite element for the Stokes equations*, Calcolo, **21**, 4 (1984), pp. 337–344.

[3] Babuška, I., *The finite element method with Lagrangian multipliers*, Numer. Math., **20** (1973), pp. 179-192.

[4] Baia, M. and Sequeira, A., *A finite element approximation for the steady solution of a second-grade fluid model*, J. Comp. and Appl. Math., **111** (1999), pp. 281–295.

[5] Bercovier, M. and Pironneau O., *Error estimates for finite element method solution of the Stokes problem in the primitive variables*, Numer. Math. **33**, (1979), pp. 211–224.

[6] Bernard, J. M., *Fluides de Second et Troisième Grade en Dimension trois: Solution Globale et Régularité*, Thesis University Pierre et Marie Curie, Paris VI, 1998.

[7] Bernard, J. M., *Stationary problem of second-grade fluids in three dimensions: existence, uniqueness and regularity*, Math. Meth. Appl. Sci., **22** (1999), pp. 655–687.

[8] Bernardi, C., *Optimal finite element interpolation on curved domains*, SIAM J. Numer. Anal., **26**, (1989), pp. 1212–1240.

[9] Bernardi, C. and Girault, V., *A local regularization operator for triangular and quadrilateral finite elements*, SIAM J. Numer. Anal., **35**, 5 (1998), pp. 1893–1916.

[10] Boland, J. and Nicolaides, R., *Stablility of finite elements under divergence constraints*, SIAM J. Numer. Anal., **20**, 4 (1983), pp. 722–731.

[11] Brenner, S. and Scott, L. R., *The Mathematical Theory of Finite Element Methods*, TAM **15**, Springer-Verlag, Berlin, 1994.

[12] Bresch, D. and Lemoine, J., *Stationary solutions for second-grade fluids equations*, M3AS, **8**, (1998).

[13] Brezzi, F., *On the existence, uniqueness and approximation of saddle-point problems arising from Lagrange multipliers*, RAIRO, Anal. Num., **R2** (1974), pp. 129-151.

[14] Brezzi, F. and Falk, R. S., *Stability of a higher order Hood-Taylor method*, SIAM J. Numer. Anal., **28**, (1991), pp. 581–590.

[15] Brezzi, F. and Fortin, M., *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.

[16] Ciarlet, P. G., *Basic error estimates for elliptic problems - Finite Element Methods, Part 1*, Handbook of Numerical Analysis, P. G. Ciarlet and J. L. Lions, eds., North-Holland, Amsterdam, 1991.

[17] Cioranescu, D. and Ouazar, E. H., *Existence et unicité pour les fluides de second grade*, Note CRAS **298**, Série I (1984), pp. 285–287.

[18] Cioranescu, D. and Ouazar, E. H., *Existence and uniqueness for fluids of second grade*, in *Nonlinear Partial Differential Equations, Collège de France Seminar*, Pitman **109**, Boston, MA (1984), pp. 178–197.

[19] Cioranescu, D. and Girault, V., *Weak and classical solutions of a family of second grade fluids*, Int. J. Non-Linear Mech. **32** (1997), pp. 317–335.

[20] Clément, P., *Approximation by finite element functions using local regularization*, RAIRO, Anal. Num. **R-2** (1975), pp. 77–84.

[21] DiPerna, R. J. and Lions, P. L., *Ordinary differential equations, transport theory and Sobolev spaces*, Invent. math. **98** (1989), pp. 511–547.

[22] Dunn, J. E. and Fosdick, R. L., *Thermodynamics, stability, and boundedness of fluids of complexity two and fluids of second grade*, Arch. Rational Mech. Anal. **56**, 3 (1974), pp. 191–252.

[23] Dunn, J. E. and Rajagopal, K. R., *Fluids of differential type: Critical review and thermodynamic analysis*, Int. J. Engng Sci. **33**, 5 (1995), pp. 689–729.

[24] Durán, R., Nochetto, R. H. and Wang, J., *Sharp maximum norm error estimates for finite element approximations of the Stokes problem in* $2 - D$, Math. Comp. **51**, 184 (1988), pp. 1177–1192.

[25] Fortin, M., *An analysis of the convergence of mixed finite element methods*, RAIRO Anal. Numér., **11**, R3 (1977), pp. 341–354.

[26] Galdi, G. P., Grobelaar-Van Dalsen, M. and Sauer, N., *Existence and uniqueness of classical solutions of the equations of motion for second grade fluids*, Arch. Rat. Mech. Anal. **124**, (1993), pp. 221-237.

[27] Galdi, G. P. and Sequeira, A., *Further existence results for classical solutions of the equations of second grade fluids*, Arch. Rat. Mech. Anal. **128**, (1994), pp. 297-312.

[28] Girault, V., *A local projection operator for quadrilateral finite elements*, Math. of Comp., **64**, (1995), pp. 1421–1431.

[29] Girault, V. and Raviart, P. A., *Finite Element Methods for Navier-Stokes Equations. Theory and Algorithms*, SCM **5**, Springer-Verlag, Berlin, 1986.

[30] Girault, V. and Scott, L. R., *Analysis of a two-dimensional grade-two fluid model with a tangential boundary condition*, J. Math. Pures Appl. **78**, 10 (1999), pp. 981–1011.

[31] Girault, V. and Scott, L. R., *Finite-element discretizations of a two-dimensional grade-two fluid model*, M2AN **35**, (2002), pp. 1007–1053.

[32] Girault, V. and Scott, L. R., *Hermite interpolation of non-smooth functions preserving boundary conditions*, Math. of Comp. **71**, (2002), pp. 1043–1074.

[33] Girault, V. and Scott, L. R., *Upwind discretizations of a steady grade-two fluid model in two dimensions*, in *Studies in Mathematics and its Applications*, **31**, (2002), pp. 393–414.

[34] Girault, V. and Scott, L. R., *A quasi-local interpolation operator preserving the discrete divergence*, Calcolo, **40** (2003), pp. 1–19.

[35] Grisvard, P., *Elliptic Problems in Nonsmooth Domains*, Pitman Monographs and Studies in Mathematics, **24**, Pitman, Boston, MA, 1985.

[36] Holm, D. D., Marsden, J. E. and Ratiu, T. S., *Euler-Poincaré models of ideal fluids with nonlinear dispersion*, Phys. Rev. Lett. **349** (1998), pp. 4173–4177.

[37] Holm, D. D., Marsden, J. E. and Ratiu, T. S., *The Euler-Poincaré equations and semidirect products with applications to continuum theories*, Adv. in Math. **137** (1998), pp. 1–81.

[38] Hood, P. and Taylor, C., *A numerical solution of the Navier-Stokes equations using the finite element technique*, Comp. and Fluids 1, (1973), pp. 73–100.

[39] Hopf, E., *Uber die Aufangswertaufgabe fur die hydrodynamischen Grundleichungen*, Math. Nachr. **4**, (1951), pp. 213–231.

[40] Hugues, T. J. R., *A simple finite element scheme for developping upwind finite elements*, Int. J. Numer. Meth. Eng. **12** (1978), pp. 1359–1365.

[41] Johnson, C., *Numerical solution of PDE by the finite element method*, Cambridge University Press, Cambridge, 1987.

[42] Leray, J., *Etude de diverses équations intégrales nonlinéaires et de quelques problèmes que pose l'hydrodynamique*, J. Math. Pures Appl. **12** (1933), pp. 1–82.

[43] Lesaint, P. and Raviart, P. A., *On a finite element method for solving the neutron transport equation*, in *Mathematical Aspects of finite Elements in Partial Differential Equations*, pp. 89–122, Academic Press, New York, NY, 1974.

[44] Lions, J. L., *Quelques Méthodes de Résolution des Problèmes aux Limites Non Linéaires*, Dunod, Paris, 1969.

[45] Lions, J. L. and Magenes, E., *Problèmes aux Limites non Homogènes et Applications, I*, Dunod, Paris, 1968.

[46] Nečas, J., *Les Méthodes Directes en Théorie des Équations Elliptiques*, Masson, Paris, 1967.

[47] Ortega, R. R., *Contribución al estudio teórico de algunas E.D.P. no lineales relacionadas con fluidos no Newtonianos*, Thesis University of Sevilla, Spain, 1995.

[48] Ouazar, E. H., *Sur les Fluides de Second Grade*, Thesis University Pierre et Marie Curie, Paris VI, 1981.

[49] Park, K. H., *Least-Squares Methods for the Simulation of the Flow of an Incompressible Fluid of Second Grade*, Ph.D. Thesis University of Houston, 1998.

[50] Peetre, J., *Espaces d'interpolation et théorème de Soboleff*, Ann. Inst. Fourier **16** (1966), pp. 279–317.

[51] Picasso, M. and Rappaz, J., *Existence, a priori and a posteriori error estimates for a nonlinear three fields Stokes problem arising from viscoelastic flows*, M2AN, **35**, (2001), pp. 879–897.

[52] Pironneau, O., *Finite Element Methods for Fluids*, Wiley, 1989.

[53] Puel, J. P. and Roptin, M. C., *Lemme de Friedrichs. Théorème de densité résultant du lemme de Friedrichs*, work supervised by C. Goulaouic, Diplôme d'Etudes Approfondies, University of Rennes, France, 1967.

[54] Rajagopal, K. R., *On boundary conditions for fluids of the differential type*, dans *Navier-Stokes Equations and Related Problems*, Plenum Press, 1995.

[55] Rivlin, R. S. and Ericksen, J. L., *Stress-deformation relations for isotropic materials*, Arch. Rational Mech. Anal. **4** (1955), pp. 323–425.

[56] Scott, L. R. and Zhang, S., *Finite element interpolation of non-smooth functions satisfying boundary conditions*, Math. Comp., **54** (1990), pp. 483–493.

[57] Stein, E., *Singular integrals and differentiability properties of functions*, Princeton University Press, Princeton, NJ, 1970.

[58] Stenberg, R., *Analysis of finite element methods for the Stokes problem: a unified approach*, Math. Comp. **42**, (1984), pp. 9–23.

[59] Tartar, L., *Topics in Nonlinear Analysis*, Publications Mathématiques d'Orsay, Université Paris-Sud, Orsay, 1978.

[60] Temam, R., *Navier-Stokes Equations, Theory and Numerical Analysis*, North-Holland, Amsterdam, 1979.

[61] Temam, R., *Infinite-dimensional dynamical systems in mechanics and physics*, Applied Mathematical Sciences, **68**, Springer-Verlag, Berlin, 1997.

[62] Verfürth, R., *Error estimates for a mixed finite element approximation of the Stokes equations*, RAIRO Anal. Numér. **18**, 2, (1984), pp. 175–182.

[63] Videmann, J. H., *Mathematical analysis of visco-elastic non-newtonian fluids*, Thesis, University of Lisbon, 1997.

Shape Optimisation

P. Le Tallec*, M. Halard, E. Laporte,
Ecole Polytechnique
91 128 Palaiseau Cedex, France
e-mail: patrick.letallec@polytechnique.fr

Abstract The course presents the basic tools for handling shape optimisation problems: formulation, discretisation and reduction to a mathematical programming problem, mesh deformation, optimisation by interior point techniques, fast calculation of gradients by adjoint state techniques and automatic software differentiation. It also discusses possible improvement techniques such as mesh independent strategies and one shot methods.

Keywords: *shape optimisation, mesh adaption, interior point techniques, one shot methods.*

1 Introduction

The present course presents the numerical methodology which can be efficiently used in optimal design. The considered point of view is local. It assumes that a reference shape has been decided and that key design constraints have been identified. The systems under consideration are those used in computational structural or fluid mechanics and are modelled by large scale partial differential equations.

The mathematical framework of such problems has been introduced thirty years ago in [3, 22]. Its practical implementation has been spectacularly improved in the recent years due to progress in discretisation strategies, optimisation algorithms, or automatic differentiation [19]. Indeed a practical optimisation strategy combines different tools which have all been subject to recent developments [13, 18]:

- an adequate discretisation algorithm of the governing differential equations reducing them to large scale finite dimensional models [15, 25]
- specific liftings to smoothly deform any admissible reference configuration and its corresponding volumic discretisation grid or finite element mesh [8];
- efficient optimisation algorithms respecting design constraints [9, 2]
- fast calculation of gradients by adjoint state techniques and automatic software differentiation [17, 18]

The methodology described herein will cover these different aspects, and in particular the control of the shape's deformation by spline interpolation, mesh transformation strategies, and constrained optimization introducing the (Euler Lagrange) optimality conditions and solving them by interior point techniques. Several developments are then introduced in order to improve mesh independance, robustness and efficiency. Most of the material covered is taken from [13], where further details can be found together with available softwares (see also [14, 12].

2 Problem's formulation

2.1 A steady problem in fluid dynamics

A generic problem in aerodynamics is to minimise the drag of an airfoil with respect to its shape γ under some given constraints (cf. fig. 1). The constraints can be either geometric (volume, ...) or aerodynamic

(lift, ...). The flow around the airfoil is characterised by the state variables $W_\gamma(t, \vec{x})$, typically density

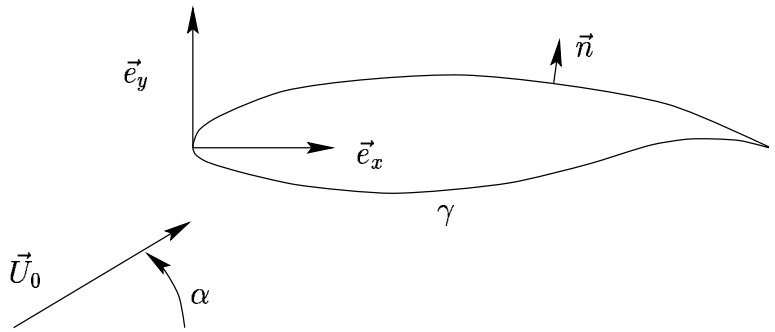


Figure 1: Airfoil in an unbounded domain ($\vec{U}_0$ = free stream velocity)

velocity and energy, defined for all $\vec{x}$ outside γ and satisfying the state equation

$$\frac{\partial W_\gamma}{\partial t} = E(t, \gamma, W_\gamma), \tag{1}$$

where E can be the Euler or the Navier-Stokes equations. For periodic problems of period T, we add the following conditions:

$$W_\gamma(0, \vec{x}) = W_\gamma(T, \vec{x}) \quad \forall \vec{x},$$

and we use cost functions of the form

$$j(\gamma) = \int_0^T J(\gamma, W_\gamma(t))\, dt. \tag{2}$$

For example, J can represent the drag around the body γ in a flow characterised by the state W_γ. In steady cases, the state equation reduces to

$$E(\gamma, W_\gamma) = 0, \tag{3}$$

and the cost function reduces to

$$j(\gamma) = J(\gamma, W_\gamma). \tag{4}$$

The constraints can be mechanical functions $g_1(\gamma, W_\gamma)$ like the drag, or geometric ones like minimal volume $g_2(\gamma)$ or imposed positions $h_1(\gamma)$ of specific points.

As seen later, the shape γ will be defined by a set of parameters

$$\alpha = (\alpha_i)_{i=1,\dots,n} \in \mathbb{R}^n,$$

with $n \in \mathbb{N}$. The final problem then writes

$$\min_{\substack{\alpha \in \mathbb{R}^n \\ g(\alpha, W_\gamma) \leq 0 \\ h(\alpha, W_\gamma) = 0}} j(\alpha) = J(\alpha, W_\gamma(\alpha)), \tag{5}$$

the state $W_\gamma(\alpha)$ being obtained by solving the relevant state equation

$$E(\gamma(\alpha), W_\gamma) = 0. \tag{6}$$

2.2 Shape optimisation for nonlinear structures

A basic structural shape optimisation problem was proposed in [6]. The state equation describes the equilibrium of a nonlinear structure made of a compressible hyperelastic material in large displacements. The structure under consideration is a thick beam of length $L = 1m$, width $l = 0.3m$ and heigth $h = 0.20m$, clamped on one end, and subjected at the other end to an imposed vertical displacement, combined with zero axial displacement and shear forces. Within the structure, the constitutive law is of Saint-Venant Kirchhoff type, associated to a free energy density $\psi(E)$ given in function of the Green Lagrange strain tensor $\underline{\underline{E}} = \frac{1}{2}\left(\underline{\underline{F}}^T\underline{\underline{F}} - \underline{\underline{Id}}\right)$ by

$$\psi(\underline{\underline{E}}) = \frac{\lambda_E}{2}(Tr\underline{\underline{E}})^2 + \mu_E Tr(\underline{\underline{E}}^2),$$

with Lamé coefficients $\lambda_E = 0.35\,Gpa$ and $\mu_E = 0.7\,Gpa$. The displacement field $W_\gamma = \underline{\xi}$ inside this structure, directly related to the deformation tensor $\underline{\underline{F}} = \underline{grad}(\underline{x} + \underline{\xi})$ is the solution of the equilibrium equations given in variational form by

$$E\left(\gamma(\alpha), W_\gamma\right) := \int_\Omega \underline{\underline{F}}^T \cdot \frac{\partial \psi}{\partial \underline{\underline{E}}} : \underline{grad}\left(\underline{\hat{v}}\right) - \int_\Omega \underline{f} \cdot \underline{\hat{v}} = 0, \forall \underline{\hat{v}} \in V_\Omega, \tag{7}$$

with

$$V_\Omega = \{\underline{v} : \Omega \to \mathbb{R}^3, \underline{v}|_{\partial\Omega_{int}} = 0, \underline{v} \in H^1(\Omega)\}$$

the space of kinematically admissible velocity fields.

The control parameters are the vertical positions of the four vertices of the mid cross section (see figure 2), measured by difference with the positions of the corresponding vertices at the end sections. In other words, the choice $\alpha = (r_1, r_2, r_3, r_4) = (0, 0, 0, 0)$ corresponds to a straight beam.

Denoting by ξ_z^0 the vertical displacement observed in a given reference design, the optimal design problem now consists in approximating at best a given displacement field $\xi_z^0 + 0,05m$ on the top surface $\Gamma_{sup}(z)$ of the beam. For this purpose, we define the cost function by

$$\begin{aligned} &\alpha = (r_1, r_2, r_3, r_4), \\ &W_\gamma = \underline{\xi}, \\ &J(\alpha, W_\gamma) = \int_{\Gamma_{sup}(z)} (\xi_z - \xi_z^0 - 0,05)^2 da. \end{aligned} \tag{8}$$

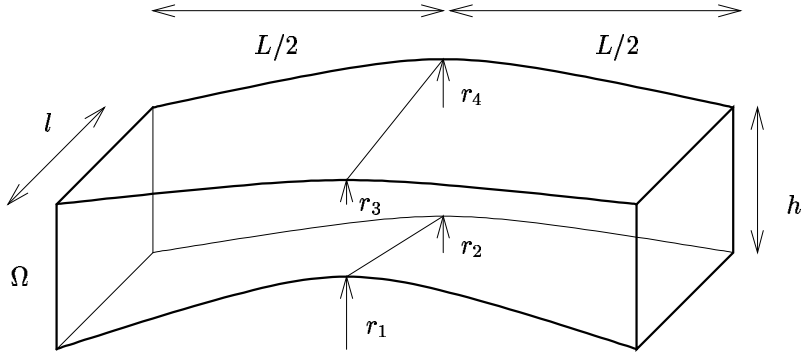
The final problem is again given by (5), with the cost function and state equation being given by (8) and (7), respectively.

3 Gradient based optimisation strategy

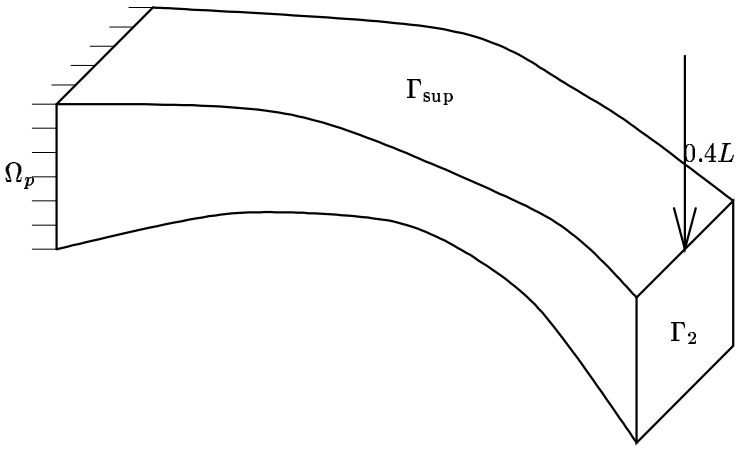
3.1 Discretisation

The above optimisation problems can be fully described through the numerical solvers to be used during its solution. The constitutive ingredients are then

1. a reference configuration $\Omega^0 \in \Omega_{adm}$, and the corresponding contour γ^0;
2. a map $\gamma\left(\alpha, \gamma^0\right)$ defining the shape γ of the contour from its initial shape and from the normal motion $(\alpha_i)_{i=1,n}$ of carefully chosen control points (see for example section 3.3);
3. a computational grid $\{X(\alpha) = (x_i)_{i=1,ns}, NU = (NU_{i,l})_{i=1,ndloc,l=1,nt}\}$ defined on the configuration Ω delimited by the contour γ describing the position $X \in \mathbb{R}^3$ of the ns grid nodes, and giving for each cell $l = 1$ to NT the set $NU_{.,l}$ of its $ndloc$ vertices. In this construction, the shape of the object is entirely characterised by its computational grid, and the parameters α controlling the



Design parameters : vertical displacements of the mid section vertices.



Imposed displacement on the end section Γ_2.

Figure 2: The model mechanical problem : optimal design of a clamped beam in large deformations. Taken from [6].

shape are only used to control the grid deformation. For example, in the structure introduced in the previous section, the coordinates $\underline{X}(z)$ of the finite element nodes are given explicitly in terms of the design parameters α by parabolic interpolation between the end and mid sections, the calculations being performed on regular structured meshes, with a number NT of second order hexaedral finite elements ranging from $NT = 48$ to $NT = 960$ (Fig 16).

4. the discretisation scheme

$$E_h(X, W_h) = 0 \text{ in } \mathbb{R}^q$$

used for approximating the state equation $E = 0$ on the computational grid X, and for defining the approximate discrete solution $W_h \in \mathbb{R}^q$ of the state equation.

With this description, the design constraints $g(\gamma) \leq 0$ can be also expressed as functions of the present position of the grid points X and can be reduced to the generic form

$$g_{h_i}(\alpha, X) \leq 0, \forall 1 \leq i \leq m.$$

Altogether, the shape opitmisation problem is then reduced to the approximate finite dimensional problem

$$(P_h) \min_{\alpha \in \mathbb{R}^n, g_h(\alpha, X) \leq 0, h(\alpha, X) = 0} j_h(\alpha) = J(X(\alpha), W_h(\alpha)),$$

where $(X(\alpha), W_h(\alpha))$ denotes the solution of the discrete state equation

$$(E_h) \begin{cases} E_h(X, W_h) = 0 \text{ in } \mathbb{R}^p, \\ X = X(\alpha). \end{cases}$$

This new formulation (P_h) is convenient for two main reasons:

- it has the form of a classical finite dimensional constrained optimisation problem to be solved by standard techniques of mathematical programming;
- all functions appearing in this formulation can be explicitly calculated on a computer.

On the other hand, this formulation ignores the natural topology of the problem under study and is dependent on the quality of the discretisation strategy which is used. This can be overcome by using adaptive discretisation techniques, as explained in [8]. The generic difficulty consists then in the construction and differentiation of the map $\alpha \mapsto X(\alpha)$ describing the evolution of the grid as a function of the control parameters α. This construction is described for example in [12] and in section 3.4.

3.2 Basic minimisation strategy

After discretisation, the above optimisation problems can be solved by any standard gradient based algorithm. The corresponding flowchart is:

1. let $\alpha^k = (\alpha_{k,i})_{i=1,\ldots,n} \in \mathbb{R}^n$ be the current value of the design parameters;
2. compute the gradient $\frac{dj_h}{dX}$ of the cost function with respect to the coordinates using for example Lagrangian techniques:
$$\frac{dj_h}{dX} = \frac{\partial L}{\partial X}(X, W, P^*),$$
where W denotes the full vector of the state variables of the problem (aerodynamic and mechanic), P^* the corresponding vector of the adjoint variables and $L = J(X, W) + \langle P, E_h(X, W) \rangle$ the Lagrangian of the optimisation problem, the state equation (E_h) being considered in this gradient calculation as an imposed equality constraint;
3. compute the gradient $\frac{dj_h}{d\alpha} = \frac{dj_h}{dX}\frac{dX}{d\alpha}$ of the cost function with respect to the design parameters;
4. update α by $\alpha^{k+1} = \alpha^k - S_k^{-1}\frac{dj_h}{d\alpha}$.

Above S_k is any convenient definite positive approximation of the Hessian of j (cf. for example [10]).

When introducing design constraints $h = 0, g \leq 0$, this simple minded Newton's strategy should be replaced by more efficient tools such as interior point techniques solving the Euler Lagrange optimality conditions of Problem (P_h) [1, 24]

$$\nabla j(\alpha^*) + \sum_{i=1}^{q} \mu_i^* \nabla h_i(\alpha^*) + \sum_{l=1}^{m} \lambda_l^* \nabla g_l(\alpha^*) = 0 \text{ in } \mathbb{R}^n, \tag{9}$$

$$h_i(\alpha^*) = 0, \forall i = 1, q, \tag{10}$$

$$\lambda_l^* g_l(\alpha^*) = 0, \forall l = 1, m, \tag{11}$$

$$(\alpha^*, \mu^*, \lambda^*) \in C = \{(\alpha, \mu, \lambda) \in \mathbb{R}^n \times \mathbb{R}^q \times \mathbb{R}^m, g(\alpha) \leq 0, \lambda \geq 0\}. \tag{12}$$

Indeed, for general constraint functions g_l, the direction $d\alpha = S_k^{-1}\frac{dj_h}{d\alpha}$ of update proposed by the Newton step may point to the outside of the admissible domain C. It is then be impossible to find any

admissible update on this direction and the algorithm fails. Moreover, a global analysis of convergence indicates that the most efficient algorithms are obtained when the different updates (α^k, λ^k) approach all active boundaries of C at the same speed, that is when all products $\lambda_l^k g_l(\alpha^k)$ stay close to a given average value $\omega_k \approx C\|dz\|$. These two arguments make it necessary to modify the Newton direction of search in order to deflect it towards the center of C. Because of this, an interior point iteration has three main steps :

1) Newton iteration solving the full linearised optimality system in $(d\alpha, \lambda, \mu)$

$$
\begin{aligned}
&B \cdot d\alpha + \sum_{i=1}^{q} \mu_i \nabla h_i(\alpha^k) + \sum_{l=1}^{m} \lambda_l \nabla g_l(\alpha^k) = -\nabla j(\alpha^k), \\
&\nabla h_i(\alpha^k) \cdot d\alpha = -h_i(\alpha^k), \forall i = 1, q, \\
&\lambda_l^k \nabla g_l(\alpha^k) \cdot d\alpha + \lambda_l g_l(\alpha^k) = 0, \forall l = 1, m,
\end{aligned}
$$

the matrix B being an adequate approximation of the full Hessian

$$
\nabla^2 j(\alpha^k) + \sum_{i=1}^{q} \mu_i^k \nabla^2 h_i(\alpha^k) + \sum_{l=1}^{m} \lambda_l^k \nabla^2 g_l(\alpha^k).
$$

2) Deflexion step building an updated direction $d\alpha + \rho\tilde{d}$ which is a strict direction of descent for the cost j but which points towards the inside of domain C. This update is obtained by solving the same Newton system as above, but with a new right hand side made of m strictly positive numbers $(\omega_l)_{l=1,m}$

$$
\begin{aligned}
&B \cdot \tilde{d} + \sum_{i=1}^{q} \tilde{\mu}_i \nabla h_i(\alpha^k) + \sum_{l=1}^{m} \tilde{\lambda}_l \nabla g_l(\alpha^k) = 0, \\
&\nabla h_i(\alpha^k) \cdot \tilde{d} = 0, \forall i = 1, q, \\
&\lambda_l^k \nabla g_l(\alpha^k) \cdot \tilde{d} + \tilde{\lambda}_l g_l(\alpha^k) = -\omega_l, \forall l = 1, m.
\end{aligned}
$$

By construction, the proposed right hand side forces the products $\lambda_l g_l$ towards negative values. The direction of deflexion $\tilde{d}$ is then a linear combination of the constraints gradients with strictly negative coefficients (Figure 3), and therefore points towards the interior of C.

3) Line search with scalar unknown t minimizing a penalized cost function j_c on the line of descent $\alpha^k + t(d\alpha + \rho\tilde{d}) \in C$

$$
\min_{\substack{t \in \mathbb{R} \\ \alpha^k + t(d\alpha + \rho\tilde{d}) \in C}} j_c(\alpha^k + t(d\alpha + \rho\tilde{d})).
$$

3.3 Parametrisation of a shape deformation.

The next problem is to update a shape as characterised by the data of consecutive boundary vertices $M_1, M_2, \ldots$, which are either the control points used in the spline construction, or boundary vertices of an existing computational grid. For two dimensional contours, at each vertex M_i, the user is able to construct a vector $\underline{n}_i$ approximatively perpendicular to the contour at this point and an approximate curvilinar abscissa s_i of this point along the contour.

The deformation of this contour can then be characterised by the displacement of each vertex M_i along the normal $\underline{n}_i$, this normal displacement being itself parametrised by a cubic spline $\psi(t)$ defined on n user defined abscissae $s_1 = t_1 < t_2 < \ldots < t_n = s_q$

$$
\psi(t) = \sum_{i=1}^{n-1} \sum_{j=1}^{4} c_{j,i}(\alpha) f_{j,i}(t)
$$

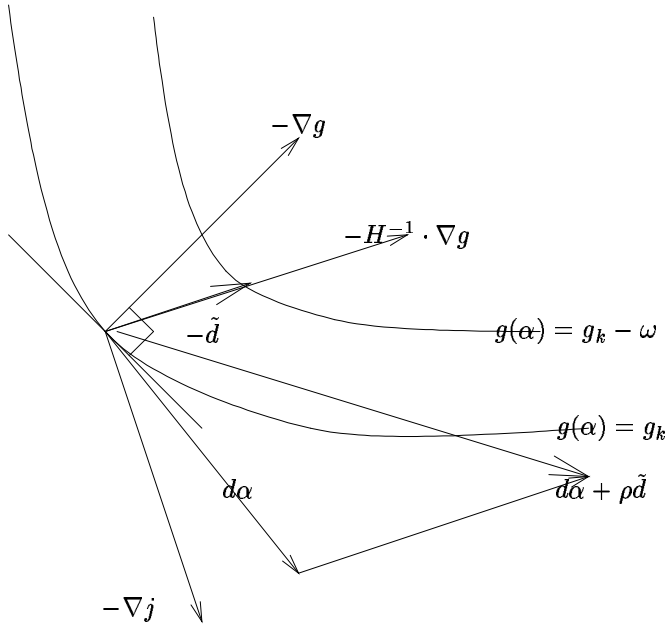


Figure 3: Unfeasibility of the Newton's direction of descent $d\alpha$ for convex constraints and construction of a deflected direction.

whose coefficients $c_{j,i}$ linearly and explicitely depend on the real parameters $(\alpha_i)_{i=1,n} = \psi'(t_1), \psi(t_i), i = 2, n-1, \psi'(t_n)$. The motion $\underline{\xi}_l$ of each boundary point of abscissa t is then given by

$$\underline{\xi}(\underline{x}(t)) = \psi(t)\underline{n}(t) = \left(\sum_{i=1}^{n-1} \sum_{j=1}^{4} c_{j,i}(\alpha) f_{j,i}(t) \right) \underline{n}(t), \tag{13}$$

building the shape $\gamma = \underline{x} + \underline{\xi}$ of the updated contour. Different generalisations are possible for the description and updating surfaces in three dimensions. The simplest one, traditional for example in naval architecture, consists in interpolating cross contours characterized by splines. More elaborate G^1 type interpolations are also possible.

3.4 Mesh updating

Generating a mesh of Ω is usually based on the existence of a preexisting mesh $X^0 = \left(\left(x_i^0 \right), NU^0 \right)$ of the underlying reference configuration. The map γ defines then the motion of the boundary grid points $(x_i)_{i \in \gamma}$. Two strategies can then be followed to update the internal grid points : the first is to use an explicit and differentiable lifting operator propagating the motion of the boundary nodes to the inside nodes through an adequate averaging strategy. For example, if we write

- $\underline{\delta}_k$ the motion of the boundary node x_k^0;
- w_k given coefficents associated each boundary node $x_k^0 \in \gamma^0$;
- $\beta \geq 2$ an arbitrary coefficient;
- for each internal node $x_i^0 \notin \gamma^0$,

$$c_{ki} = \frac{1}{|x_k^0 - x_i^0|^\beta},$$

and

$$c_i = \sum_{x_k^0 \in \gamma^0} w_k c_{ki},$$

then the operator $\mathcal{R}_h$ defined for each $x_i^0 \notin \gamma^0$ by

$$\mathcal{R}_h\left(x_i^0, \underline{\delta}\right) = \frac{1}{c_i} \sum_{x_k^0 \in \gamma^0} w_k c_{ki} \underline{\delta}_k$$

can be used for this purpose. This operator is robust but time consuming (computational time proportional to the total number of nodes times the number of boundary nodes). Another lifting operator, less time consuming but less robust, is given by the following sequence of operations :

1. the motion of all the internal nodes is set to zero;
2. for each grid cell or each finite element, compute the average motion of the nodes of the cell or of the element;
3. for each internal node, compute the average motion of the cells or of elements which the node belongs to;
4. iterate steps 2 and 3.

The second strategy to generate a mesh of Ω is adaptative remeshing where nodes are added or removed, and edges are swapped in order to fulfill an accuracy requirement defined simultaneously on the state equation and on the adjoint equation (cf. for example [8]). This choice is much more robust, but cannot reduce the mesh updating to a differentiable map $X = X(\alpha)$.

3.5 Numerical calculation of the gradient of the cost function.

The lagrangian approach leads to a very general numerical algorithm for computing rapidly the value $j(\alpha)$ and the gradient $\nabla j(\alpha)$ of the cost function. It only requires two specific numerical solvers : one for solving the state equation, one for computing the adjoint state. It is organized as follows:

1. Calculation of the grid deformation $\underline{X} = \underline{X}(\alpha)$ by the techniques of sections 3.3 and 3.4.

2. Call of the numerical solver of the state equation $E_h(\underline{X}, W_h) = 0$.

3. Call of the computer programme computing the cost function $j(\alpha) = J(\underline{X}, W_h)$.

4. Calculation of the gradients

   $$\frac{\partial J}{\partial \underline{X}} \text{ and } \frac{\partial J}{\partial W},$$

   using automatic differentiation in adjoint mode of the programme calculating the cost function $j(\alpha) = J(\underline{X}, W_h)$.

5. Solution of the adjoint system

   $$\left(\frac{\partial E_h(\underline{X}, W_h)}{\partial W}\right)^t|_{(\underline{X},W)} \cdot P = -\frac{\partial J(\underline{X}, W)}{\partial W}|_{(\underline{X},W)}.$$

6. Calculation of the cotangent directional derivative

$$\left(\frac{\partial E_h(\underline{X}, W_h(\alpha))}{\partial \underline{X}}\right)^t \cdot P(\alpha),$$

by automatic differentiation in adjoint mode of the discrete state equation $E_h(\underline{X}, W_h) = 0$.

7. Calculation of the grid gradient

$$\frac{dJ}{d\underline{X}} = \frac{\partial J(\underline{X}, W(\alpha))}{\partial \underline{X}} + \left(\frac{\partial E_h(\underline{X}, W(\alpha))}{\partial \underline{X}}\right)^t \cdot P(\alpha). \tag{14}$$

8. Final calculation of the gradient of the cost function by computing the cotangent directional derivative

$$\frac{dj}{d\alpha} = \left(\frac{\partial \underline{X}}{\partial \alpha}\right)^t \cdot \frac{dJ}{d\underline{X}},$$

using automatic differentiation in adjoint mode of the programme calculating the grid deformation $\underline{X} = \underline{X}(\alpha)$.

This algorithm can be easily adapted to the calculation of the gradient of a design constraint $g(\alpha, W(\alpha))$ if we introduce one additional adjoint state by scalar constraint.

4 Control parameters and mesh strategy

4.1 Theoretical result

The difficult part in the above algorithm is to obtain the gradient $\frac{dj}{d\alpha}$ from the gradient computed with respect to the mesh coordinates. A first choice is to use one of the smooth lifting and deformation operators $X(\alpha)$ described in section 3.4 to update the mesh from the deformation on the boundary as obtained by a given update of the control parameters α. The drawback is that no topological change in the mesh is possible. If the mesh is poor at the beginning of the optimisation process, then it will stay poor during the process, leading eventually to failure.

To overcome this problem, it was proposed in [14, 12] to use the fact that, at the analytical level, the gradient with respect to a shape does not depend on what happens inside the domain where we solve the state equation [21]. In other words, if $\theta_1 = \theta_2$ on $\partial\Omega$, then

$$\left\langle \frac{dj}{d\Omega}, \theta_1 \right\rangle = \left\langle \frac{dj}{d\Omega}, \theta_2 \right\rangle.$$

This property remains approximately valid when working at the discret level with cost j_h, at the limit where the discretisation step h goes to zero. To apply this idea, we suppose that the shape γ defining the boundary of the open set Ω_γ, to which the mesh X is associated, is parametrised by a set of parameters α

$$\alpha = (\alpha_i)_{i=1,\dots,n} \in \mathbb{R}^n,$$

with $n \in \mathbb{N}$, and that $\alpha \mapsto \gamma = \gamma(\alpha)$ is differentiable. Furthermore, we choose any given explicit lifting operator such as those described in section 3.4, e.g. a function $\mathcal{R} : H^{\frac{1}{2}}(\partial\Omega_\gamma) \mapsto H^1(\Omega_\gamma)$, that constructs from a function $g \in H^{\frac{1}{2}}(\partial\Omega_\gamma)$ a function $\bar{g} = \mathcal{R}(g) \in H^1(\Omega_\gamma)$ such that

$$\bar{g} = g \quad on \quad \partial\Omega_\gamma.$$

Then we can show that, if Ω_γ is convex, then, for $i = 1, \ldots, n$, we have

$$\left\langle \frac{dj_h}{dX}, \mathcal{R}\left(\frac{\partial \gamma}{\partial \alpha_i}\right) \right\rangle = \frac{\partial \tilde{j}}{\partial \alpha_i} + O_{h\to 0}(h),$$

where $\tilde{j}$ is the analytical cost function $\alpha \in \mathbb{R}^n \mapsto \tilde{j}(\alpha) = j\left(\Omega_{\gamma(\alpha)}\right) \in \mathbb{R}$. This means that the value $\frac{\partial j}{\partial \alpha_i}$ of the gradient does not depend much on the choice of $\mathcal{R}$. Thanks to this property, we can get in all cases a good approximation of the gradient in Ω_γ, by setting

$$\frac{\partial j}{\partial \alpha_i} = \left\langle \frac{dj}{dX}, \mathcal{R}\left(\frac{\partial \gamma}{\partial \alpha_i}\right) \right\rangle, \tag{15}$$

where $\mathcal{R}$ is any convenient lifting operator. We can therefore re-mesh or adapt the mesh when needed, or still use a smooth deformation, but we do not need to use this specific remeshing technique for calculating the gradients. The figure 5 shows the global gradient computation algorithm based on this remeshing strategy.

Remark 1 *There is another mesh independent strategy to compute gradients by using transpiration boundary conditions [4]. This technique, adapted from aeroelastic models [5] reduces the shape deformation effects to a simple change of boundary conditions, and therefore bypasses the need of mesh differentiation.*

4.2 Numerical experiments for the periodic heat equation

We want here to validate the computation of the gradient of a periodic cost function. In this example, the state equation is the T-periodic heat equation *inside* a bounded open set Ω of boundary $\gamma = \partial\Omega$: find u_Ω such that

$$\begin{cases} \frac{\partial u_\Omega}{\partial t} - \Delta u_\Omega = f(t) & \text{in } \Omega, \\ u_\Omega(t,.) = 0 & \text{on } \gamma = \partial\Omega \quad \forall t \in]0,T[, \\ u_\Omega(0,.) = u_\Omega(T,.) & \text{in } \Omega. \end{cases}$$

The analytical cost function is given by

$$j(\Omega) = \int_0^T \int_\Omega u_\Omega^2 \, d\Omega \, dt.$$

We can show [7, 20, 21, 23] that if $\partial\Omega$ is locally lipschitz, then j is differentiable, and if u_Ω is smooth enough (at least in $H^2(\Omega)$), then

$$\left\langle \frac{dj}{d\Omega}, \theta \right\rangle = \int_0^T \int_{\partial\Omega} \frac{\partial u}{\partial n}\frac{\partial v}{\partial n}(\theta.n) \, d\gamma \, dt,$$

where $\theta \in \left(W^{1,\infty}(\Omega)\right)^2$, and v is the solution of the adjoint state equation:

$$\begin{cases} -\frac{\partial v_\Omega}{\partial t} - \Delta v_\Omega = 2u(t) & \text{in } \Omega, \\ v_\Omega(t,.) = 0 & \text{on } \gamma = \partial\Omega \quad \forall t \in]0,T[, \\ v_\Omega(0,.) = v_\Omega(T,.) & \text{in } \Omega. \end{cases}$$

We can then compare the analytical gradient $\frac{dj}{d\Omega}$ and the discrete gradient $\frac{dj_h}{dX}$: for a mesh described by $X = (x_i)_i$ with $x_i = \left(x_i^1, x_i^2\right)$, if we introduce the continuous piecewiese linear function φ_i such that $\varphi_i(x_i) = 1$ and $\varphi_i(x_j) = 0$ for $j \neq i$, and

$$\xi_i^1 = \begin{pmatrix} \varphi_i \\ 0 \end{pmatrix}, \quad \xi_i^2 = \begin{pmatrix} 0 \\ \varphi_i \end{pmatrix},$$

then we have

$$\frac{\partial j_h}{\partial x_i^j} = \left\langle \frac{dj}{d\Omega}, \xi_i^j \right\rangle \quad \forall i, j.$$

On figure 6, we compared the numerical gradient and the analytical gradient for a periodic problem of period $T = 1$, on the square $\Omega =]0,1[\times]0,1[$, with

$$f\left(t, x^1, x^2\right) = \left[\frac{\omega^2}{2} + 2\pi^4\right] \sin\left(\omega t\right) \sin\left(\pi x^1\right) \sin\left(\pi x^2\right).$$

The size of the discretisation step h is $\frac{1}{20}$ on the straight edges. We can then show that, for a vertex $x_i = \left(x_i^1, x_i^2\right)$ on $\{0\} \times]0,1[$, we have for example

$$\left\langle \frac{dj}{d\Omega}, \xi_i \right\rangle = \begin{pmatrix} -\frac{h}{4}\pi^4 + \frac{\pi^2}{16h}\left[2\cos\left(2\pi x_i^2\right) - \cos\left(2\pi\left(x_i^2 + h\right)\right) - \cos\left(2\pi\left(x_i^2 - h\right)\right)\right] \\ 0 \end{pmatrix}$$

The values of the whole analytical gradient are described by the figure on the right in figure 6.

The discrete gradient on the left part of figure 6 was computed by (15) with the explicit lifting described in section 3.3.

4.3 Minimal drag problem

In the experiment described here, we start from the mesh of figure 7, the mesh of each new shape is adapted based on the mesh used for the previous shape, and the lifting is given by the simple iterative algorithm introduced in section 3.4.

The mesh adaptor is `BAMG` [8]. The metric used for the adaption is built according to the criterium of interpolation error: on the adapted mesh, the interpolation error of the choosen field should be lower than on the old mesh. The question is: what field shall we use to monitor this interpolation error? The answer given here is to compute the metric corresponding to the flow variables (density, x-velocity, y-velocity, energy) and to their adjoint variables, and then to intersect all these metrics by taking the lowest length among those required for the different variables. Indeed, if we look at the interpolation error of the gradient in the case of a Dirichlet problem, we can see that it depends both on the interpolation error on the state variable and also on the adjoint variable.

The flow equations are the steady Euler equation, the Mach number is 0.85. Parametrisation is the same as in 3.3. We have then 16 parameters. The cost function is the drag. We have 3 constraints:

- 2 inequality constraints:
  - the lift must remain greater than 95% of the initial lift;
  - the airfoil area must remain greater than 95% of the initial area;
- 1 equality constraint, which is a fixed point in the geometry. Here, the leading edge is fixed.

With this first mesh, we obtain a drag reduction greater than 50% in 5 optimisation iterations, and the constraint on the lift is satisfied. The figures 8 and 9 show the inital and the final isolines of the density. We can see that the accuracy of the computation increased during the optimisation.

Figures 10, 11, 12, 13, 14, 15 show the various meshes obtained during the optimisation. The mesh (a) is the initial guess. To avoid the failure of the linear search, the number of vertices was limited to 6000.

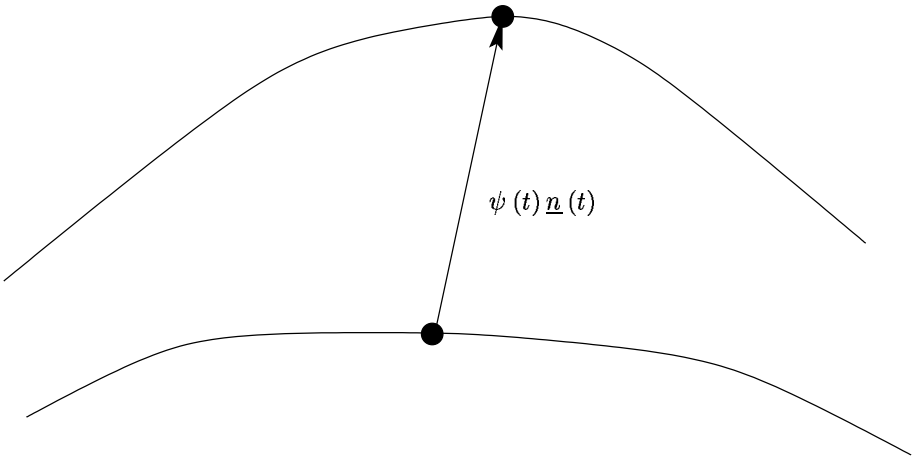


Figure 4: Motion of the boundary

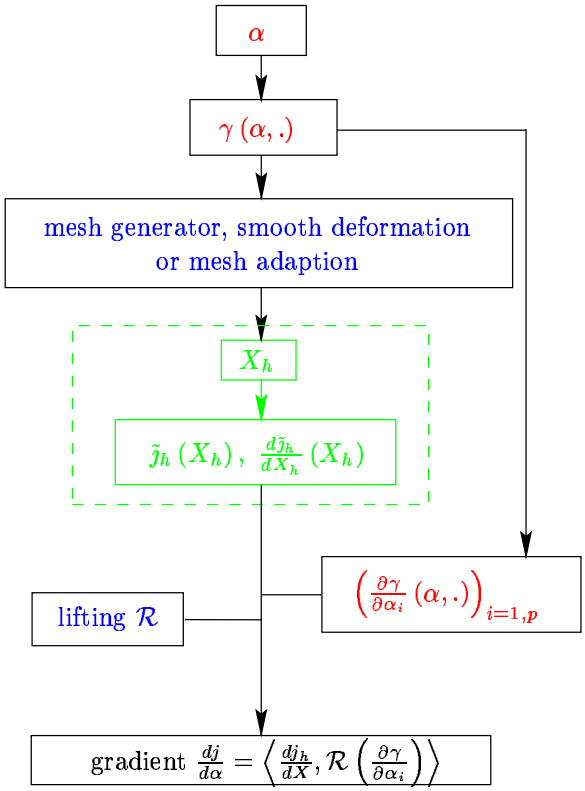


Figure 5: Global gradient computation algorithm

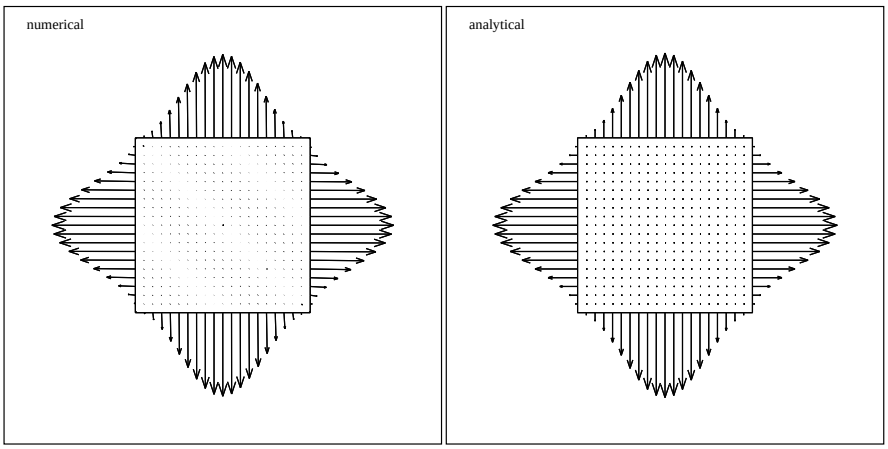


Figure 6: Comparison between numerical (left) and analytical (right) gradients

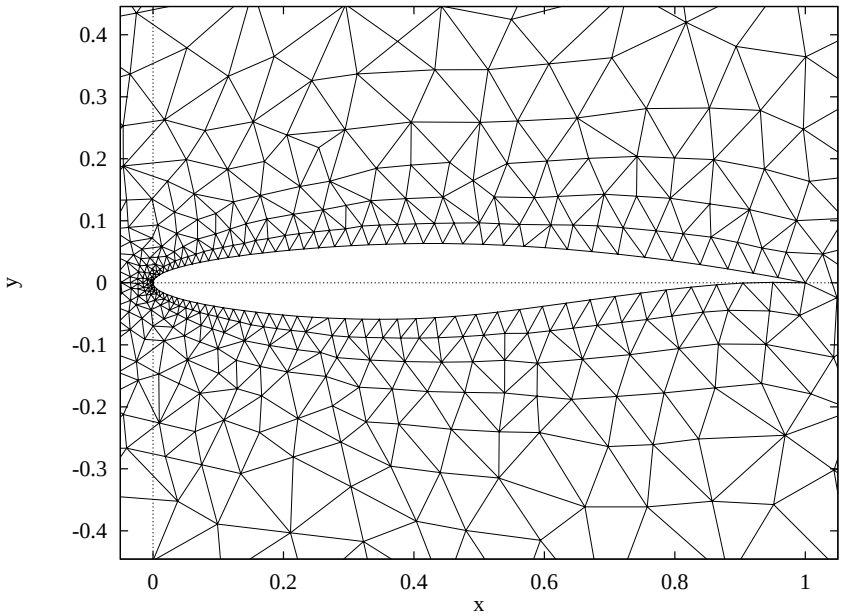


Figure 7: Initial mesh for the drag reduction problem

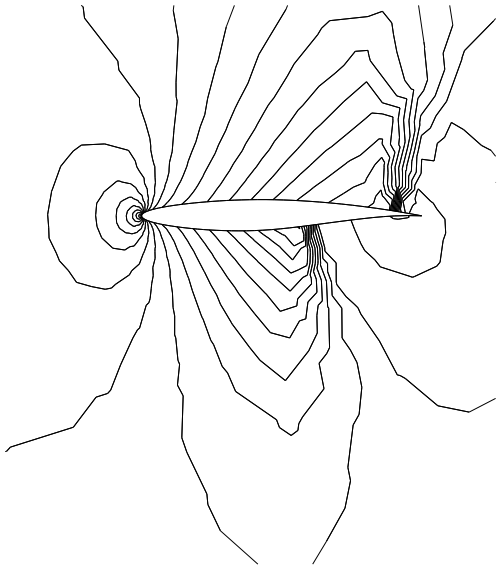


Figure 8: Iso-density lines on the reference configuration with initial mesh.

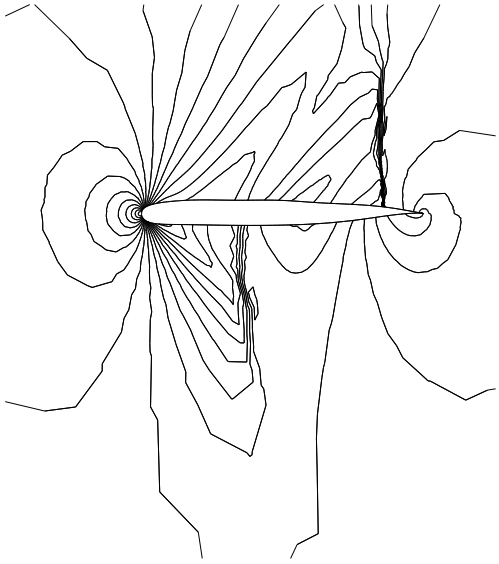


Figure 9: Iso-density lines on the optimal configuration with adapted mesh.

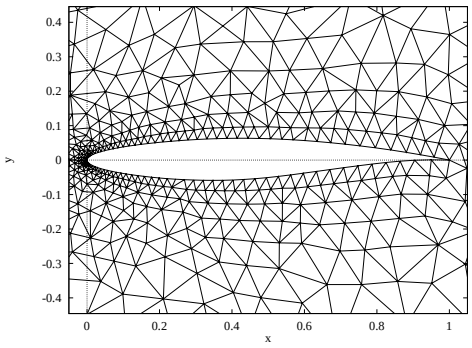


Figure 10: A first mesh with 692 vertices

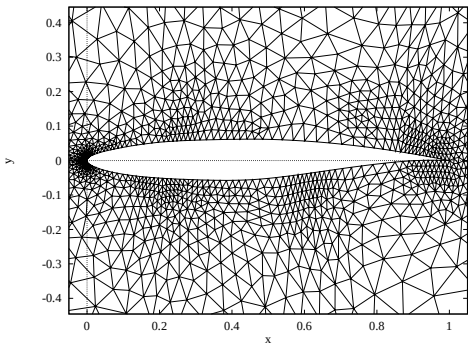


Figure 11: A second mesh with 1619 vertices

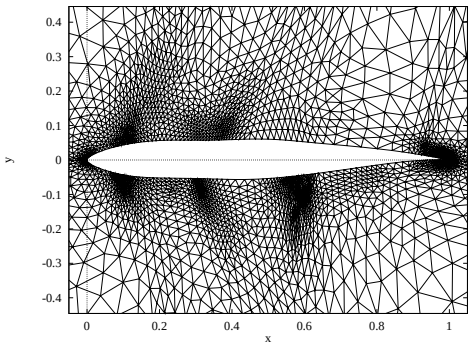


Figure 12: A third mesh with 3738 vertices

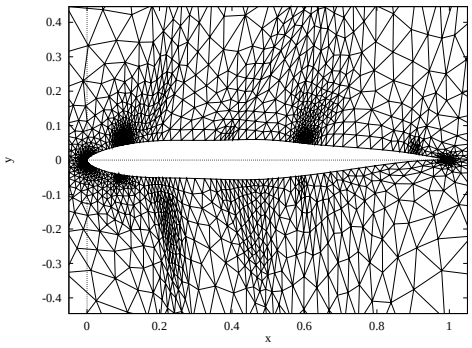


Figure 13: A fourth mesh with 2454 vertices

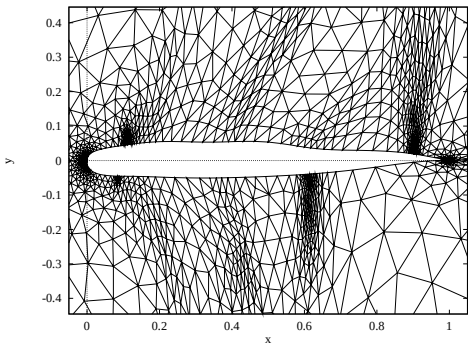


Figure 14: A fifth mesh with 1930 vertices

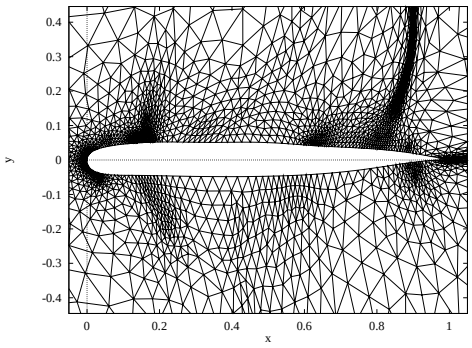


Figure 15: A sixth mesh with 3122 vertices

5 One Shot Methods

5.1 Global Algorithm

The problem in the procedure described up to now is that each evaluation of the cost function or of its gradient requires the knowledge of the state vector $W_h \in \mathbb{R}^p$, that is the solution of one occurence of the state equation in $\mathbb{R}^p$. Since p may be very large ($p \geq 10000$), and $E_h(\underline{X}, W_h)$ may be very nonlinear, solving all these occurences of the state equation is very expensive. To overcome this difficulty, one can think of relaxing the equality constraint $E_h(\underline{X}, W_h) = 0$ and add it to the optimisation programme as an external equality constraint, leading to the new problem

$$\min_{\substack{(\alpha, W_h) \in \mathbb{R}^n \times \mathbb{R}^p \\ g_l(\alpha) \leq 0 \\ E_h(\underline{X}(\alpha), W_h) = 0}} J(\underline{X}(\alpha), W_h). \tag{16}$$

After introduction of the Lagrange multipliers $\lambda \in \mathbb{R}^m$ and $\mu \in \mathbb{R}^p$ of the inequality constraints $g_l(\alpha) \leq 0$ and of the state equality constraint

$$E_h(\underline{X}(\alpha), W_h) = 0,$$

any local solution (α^*, W_h^*) of this constrained minimisation problem satisfies

$$\frac{\partial J}{\partial \alpha}(\underline{X}(\alpha^*), W_h^*) + \sum_{i=1}^{p} \mu_i^* \frac{\partial E_{hi}}{\partial \alpha}(\underline{X}(\alpha^*), W_h^*) + \sum_{l=1}^{m} \lambda_l^* \frac{\partial g_l}{\partial \alpha}(\alpha^*, W_h^*) = 0, \tag{17}$$

$$\frac{\partial J}{\partial W}(\underline{X}(\alpha^*), W_h^*) + \sum_{i=1}^{p} \mu_i^* \frac{\partial E_{hi}}{\partial W}(\underline{X}(\alpha^*), W_h^*) + \sum_{l=1}^{m} \lambda_l^* \frac{\partial g_l}{\partial W}(\alpha^*, W_h^*) = 0, \tag{18}$$

$$E_h(\underline{X}(\alpha^*), W_h^*) = 0, \tag{19}$$

$$\lambda_l g_l(\alpha^*, W_h^*) = 0, \forall l = 1, m, \tag{20}$$

$$\lambda_l \geq 0,\ g_l(\alpha^*, W_h^*) \leq 0, \forall l = 1, m. \tag{21}$$

In fluid mechanics, the above full optimality system is often solved by time marching techniques, or by similar descent methods [11]. Interior point methods turn out to be also a good method to solve this large dimensional constrained minimisation poblem (16) or its equivalent optimality conditions (17)-(21) in one single shot. These techniques apply a modified Newton's method to construct sequences of variables (α^k, W^k) that satisfy the state equation not at each step, but only at convergence : the equation of state and the optimality condition are solved simultaneously within the Newton's loop. Introducing the notation x to denote the pair (α = design, W_h = state), the diagonal matrix $\mathcal{G}$ of order m with diagonal terms g_l, the Lagrangian

$$\mathcal{L}(x, \lambda, \mu) = J(x) + \sum_{i=1}^{p} \mu_i (E_h)_i(x) + \sum_{l=1}^{m} \lambda_l g_l(x), \tag{22}$$

and a convenient approximation $B \in M_{n+p}(\mathbb{R})$ of the Hessian of the Lagrangian $\mathcal{L}$ with respect to x, we can construct a gobal matrix $H \in M_{n+2p+m}$ of the linearised problem

$$H = \begin{pmatrix} B & \frac{dE_h}{dx}^t & \frac{dg}{dx}^t \\ \frac{dE_h}{dx} & 0 & 0 \\ \frac{d\mathcal{G}}{dx}\lambda & 0 & \mathcal{G} \end{pmatrix}. \tag{23}$$

The application of an interior point algorithm to the global problem (17)-(21) leads to the following steps :

1) Newton's prediction : compute the Newton's direction by solving the linearised optimality conditions

$$H \cdot \begin{pmatrix} \delta x_o \\ \delta \mu_o \\ \delta \lambda_o \end{pmatrix} = - \begin{pmatrix} \dfrac{\partial \mathcal{L}}{\partial x}(x, \lambda, \mu) \\ E_h \\ \mathcal{G}\lambda \end{pmatrix}.$$

2) Deflexion : introduce a direction of deflexion by solving

$$H \begin{pmatrix} \delta x_1 \\ \delta \mu_1 \\ \delta \lambda_1 \end{pmatrix} = - \begin{pmatrix} 0 \\ 0 \\ (\lambda_l \omega_l)_l \end{pmatrix}$$

where ω_l is some given positive number. Once δx_1 is computed, calculate a maximum deflection step ρ such that $\delta x' = \delta x_o + \rho \delta x_1$ is a good direction of descent relatively to δx_o, in the sense :

$$\nabla J \cdot \delta x' \leq \frac{1}{2} \nabla J \cdot \delta x_o.$$

3) Line search : compute finally the next value of x (and hence of λ and μ) by minimizing J over the trajectory

$$\{x(t); x(t) = x + t\delta x', t > 0\}.$$

5.2 Reduction of the linear systems

The expensive task in the above algorithm is to solve the global very large linearised systems

$$H \begin{pmatrix} dx \\ d\mu \\ d\lambda \end{pmatrix} = r$$

of order $n + m + 2p$. If we develop the expression of H with respect to the four independent variables α, W, μ and λ, and if we assume for simplicity that the constraints g do not depend explicitely of the values of the state variables W, each of these global linear systems is of the form

$$\begin{pmatrix} B_{\alpha\alpha} & B_{\alpha W} & \frac{\partial E_h}{\partial \alpha}^t & \frac{dg}{d\alpha}^t \\ B_{W\alpha} & B_{WW} & \frac{\partial E_h}{\partial W}^t & 0 \\ \dfrac{\partial E_h}{\partial \alpha} & \frac{\partial E_h}{\partial W} & 0 & 0 \\ \dfrac{d\mathcal{G}}{d\alpha}\lambda & 0 & 0 & \mathcal{G} \end{pmatrix} \begin{pmatrix} d\alpha \\ dW \\ d\mu \\ d\lambda \end{pmatrix} = \begin{pmatrix} r_\alpha \\ r_W \\ r_\mu \\ r_\lambda \end{pmatrix}. \tag{24}$$

When the number n of control parameters is small, and if efficient solvers exist for the linearised state equation

$$\frac{\partial E_h}{\partial W} \cdot dW = r_W,$$

the above global linear system (24) can be easily solved by simple elimination of the linear increments dW and $d\mu$ of the state and adjoint state variables W and μ introducing the $p \times n$ matrix $U = (U_1, U_2, \ldots, U_n)$ obtained by solving the n linear systems of matrix $\dfrac{\partial E_h}{\partial W}(\alpha, W)$ defined by

$$\frac{\partial E_h}{\partial W} \cdot U_i = -\frac{\partial E_h}{\partial \alpha_i}, i = 1, n.$$

We can then eliminate dW and $d\mu$ from the first line of (24), and this system finally reduces to the simple linear system given by

$$\begin{pmatrix} B_{\alpha\alpha} + B_{\alpha W}U + U^t B_{W\alpha} + U^t B_{WW}U & \frac{dg}{d\alpha}^t \\ \frac{d\mathcal{G}}{d\alpha}\lambda & \mathcal{G} \end{pmatrix} \begin{pmatrix} d\alpha \\ d\lambda \end{pmatrix} = \begin{pmatrix} r_\alpha + U^t r_\mu - (B_{\alpha W} + U^t B_{WW}) \left(\frac{\partial E_h}{\partial W}\right)^{-1} \cdot r_W \\ r_\lambda \end{pmatrix}. \tag{25}$$

This linear system is of very small dimension (it is of order $n+m$) and can be solved by a residual based method such as GMRES which only requires to compute matrix vector products of the form

$$H_\alpha d\alpha_i := \Big(B_{\alpha\alpha} + B_{\alpha W}U + U^t B_{W\alpha} + U^t B_{WW}U\Big) d\alpha_i, \forall i = 1, \ldots, n,$$

which is easily computed by automatic differentiation.

5.3 Line Search

To be robust, the line search strategy must be restricted to a neighborhood of the solution curve $E_h(\underline{X}(\alpha), W_h) = 0$. For this purpose, an important enhancement consists in adding a restoration step before the global Newton's step in order to generate sequences of state variables W_h leading to small residuals in the equation of state $E_h(\underline{X}(\alpha), W_h) = 0$, resulting into a new line search procedure organized as follows

- compute first a linear update δW_0 of the state variable by solving the linearized state equation with frozen design α :
$$\left(\frac{\partial E_h}{\partial W}\right) \delta W_0 = r_W = -E_h(\alpha, W_h)\,;$$
- compute the update direction $\delta\alpha$ as predicted by the interior point algorithm when applied to the full optimality system, and compute the associated matrix of influence $U = -\left(\frac{\partial E_h}{\partial W}\right)^{-1} \cdot \frac{\partial E_h}{\partial \alpha}$;
- search for the point $(\alpha(\bar{t}), W(\bar{t})) = (\alpha + \bar{t}\delta\alpha, W_h + \delta W_0 + \bar{t}U \cdot \delta\alpha)$ on the line of search with minimal cost function.
- take as new solution the vector
$$W = W(\bar{t}) = W_h + \delta W_o + \bar{t}U \cdot \delta\alpha,$$
$$\alpha = \alpha + \bar{t}\delta\alpha.$$

5.4 Numerical Examples

We consider the case of subsection 2.2 where the right end is subjected to a large vertical displacement $\xi_z = 0,4L$ and with reference design

$$z = (r_1, r_2, r_3, r_4) = (0.01m, 0.01m, 0.02m, 0.02m).$$

The associated equilibrium solution was computed in 16 Newton iterations starting from the initial guess $W^0 = 0$, and is described in figure 16.

The problem has been solved by the above interior point algorithm with different options. Figure 17 presents a typical convergence curve, obtained in [6] when starting with $W = 0$, $z = (0.00, 0.00, 0.05, 0.01)$

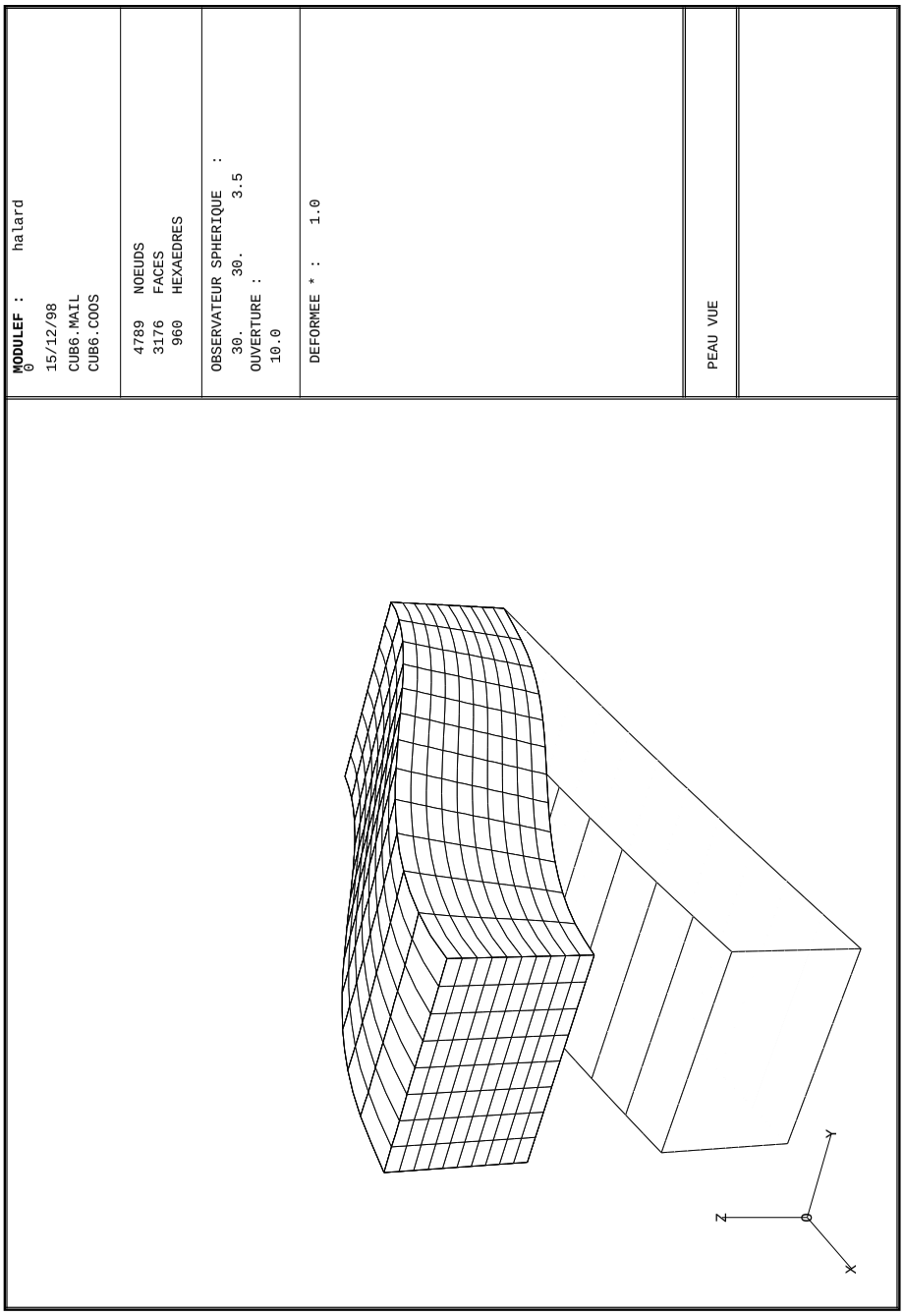


Figure 16: Initial and final shape of the reference beam. Calculation using the fine mesh with $NT = 960, p = 14367$. Taken from [6].

with an initial equilibrium recovery step using three Newton iterations. The optimal solution corresponds to $r_1 = r_2 = -0.25m$ and $r_3 = r_4 = -0.16m$. In this case, the convergence is not very sensitive to the particular line search strategy, but is more sensitive to the tolerance E_{max} used for the state equation and to a possible upper bound in the update of the control (design) parameters α. Observe that the convergence of the full optimisation loop is very similar to the convergence of the Newton's method when

applied to the solution of the state equation. Full convergence is achieved in less than 20 iterations. By observing in addition that the CPU cost of one iteration of the complete optimisation loop is only 10 percent higher than the CPU cost of a single Newton's iteration (the main cost is here the factorisation of the stiffness matrix), we can infer that in this case, finding the optimal shape is only marginally more expensive than simply computing the equilibrium solution.

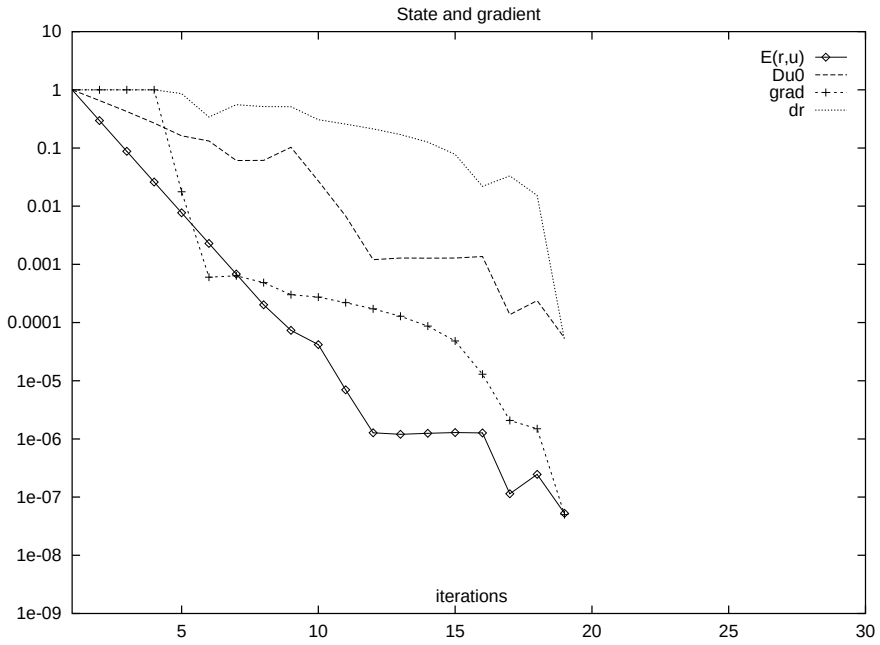


Figure 17: Convergence history of the one shot optimisation loop. Taken from [6].

6 Conclusion

The main ingredients in the optimisation methodology presented above consist in

- the reduction of the optimisation problem to a standard problem of mathematical programming in $\mathbb{R}^n$ by discretisation. This requires a correct description and discretisation of an initial shape, an efficient control of the shape's deformation by spline interpolation of the contour normal's displacement, and a good mesh deformation strategy;
- a constrained optimisation method such as interior point techniques;
- a calculation of gradients by adjoint state techniques and automatic code differentiation of the direct solvers.

We have also seen in this report that a mesh independent strategy can be used to improve robustness, without affecting too much the performance of the minimisation algorithm. Nevertheless, a lot remains to do in order to know what is the good way to control the number of vertices, the linear search and so on, to avoid the tuning of parameters and to have a good self-control of the minimisation algorithm.

Another direction of improvement was the use of a one shot methodology. This multilevel approach involves in addition a global optimization strategy of the full problem in (α, W_h), and the development of efficient algebraic solvers for the large linear systems involved. These one one methods are also of quite general use, they can be efficient and cost effective when properly used, but they are often less robust than the algorithms which operate directly in the design space $\mathbb{R}^n$.

References

[1] K.J. Arrow, L. Hurwicz, and H. Uzawa. *Studies in Nonlinear Programming.* Stanford University Press, 1958.

[2] J.F. Bonnans, J.C. Gilbert, C. Lemaréchal, and C. Sagastizabal. *Optimisation Numérique.* Springer Verlag, 1998.

[3] J. Cea. *Optimisation : théorie et algorithmes.* Dunod, Paris, 1971.

[4] M. Fernandez and M. Moubachir. Sensitivity analysis for an incompressible aeroelastic system. Rapport de Recherche RR-4264, INRIA, 2001.

[5] T. Fanion, M. Fernandez and P. Le Tallec. Deriving adequate formulations for fluid structure interactions problems : from ALE to transpiration. *Revue Européenne des Éléments Finis*, 9 (6-7):681–708, 2000.

[6] Matthieu Halard. *Conception Optimale de Formes en Elasticité Nonlinéaire.* Thèse, Université de Paris Dauphine, October 1999.

[7] J.W. He and O. Pironneau. Optimisation du profil d'aile dans un fluide visqueux. Rapport de Recherche R94015, Laboratoire d'Analyse Numérique, Université Pierre et Marie Curie, 1994.

[8] Frédéric Hecht. `BAMG`: Bidimensional anisotropic mesh generator, mars 1998.

[9] J. Herskovits. A two-stage feasible directions algorithm for nonlinear constrained optimization. *Mathematical Programming*, pages 19–38, 1986.

[10] J. Herskovits, E. Laporte, P. Le Tallec, and G. Santos. A quasi-newton interior point algorithm applied to constrained optimum design in computational fluid dynamics. *Revue Européenne des Éléments Finis*, 5(5-6):595–517, 1996.

[11] G. Kuruvila S. Ta'asan. Aerodynamic design and optimization in one-shot. *AIAA paper 92-0025*, 1992.

[12] E. Laporte. *Optimisation de formes pour écoulements instationnaires.* Thèse, Ecole Polytechnique, September 1998.

[13] E. Laporte and P. Le Tallec. *Numerical Methods in Sensitivity Analysis and Shape Optimisation.* Birkhauser, Boston, Basel, Berlin, 2002.

[14] E. Laporte and P. Le Tallec. Shape optimisation in unsteady flows. Rapport de Recherche RR-3693, INRIA, Mai 1999.

[15] B. Lucquin and O. Pironneau. *Introduction to Scientific Computing for Engineers*, Wiley, 1998.

[16] Mohamed Masmoudi. *Outils pour la Conception Optimale de Formes.* Thèse d'état, Université de Nice, 1987.

[17] B. Mohammadi. Practical applications to fluid flows of automatic differentiation for design problems. *VKI lecture*, 1997.

[18] B. Mohammadi and O. Pironneau. *Applied Optimal Shape Design for Fluids* Oxford University Press, Oxford, 2001.

[19] B. Mohammadi and O. Pironneau. New tools for optimum shape design. *CFD Review, Special Issue*, 1995.

[20] Jérome Monnier. *Optimisation de Forme pour un Système Couplé Fluide-Thermique.* Thèse de 3e cycle, Université de Nice, 1995.

[21] François Murat and Jacques Simon. Sur le contôle par un domaine géométrique. Rapport de Recherche LAN76015, Laboratoire d'Analyse Numérique, Université Pierre et Marie Curie, 1976.

[22] O. Pironneau. *Optimal shape design for elliptic systems.* Springer Verlag, 1983.

[23] Bernard Rousselet. *Quelques Résultats en Optimisation de Domaines.* Thèse d'état, Université de Nice, décembre 1982.

[24] G. Strang. *Introduction to Applied Mathematics.* Wellesley-Cambridge Press, Wellesley, Massachussets, 1986.

[25] O. Zienkiewicz. *The Finite Element Method in Engineering Science* Mc Graw-Hill, 1977.

MESHLESS ANALYSIS OF INCOMPRESSIBLE FLOWS USING THE FINITE POINT METHOD

E. Oñate, C. Sacco
International Center for Numerical Methods in Engineering (CIMNE)
Universidad Politécnica de Cataluña
Gran Capitán s/n, 08034 Barcelona, Spain
E-mail: onate@cimne.upc.es
Web page: http://www.cimne.upc.es

S. Idelsohn
Universidad Nacional del Litoral
Santa Fe, Argentina
E-mail: rnsergio@alpha.arcride.edu.ar

Abstract. A stabilized finite point method (FPM) for the meshless analysis of incompressible fluid flow problems is presented. The stabilization approach is based in the finite calculus (FIC) procedure. An enhanced fractional step procedure allowing the semi-implicit numerical solution of incompressible fluids using the FPM is described. Examples of application of the stabilized FPM to the solution of incompressible flow problems are presented.

1 INTRODUCTION

Mesh free techniques have become quite popular in computational mechanics. A family of mesh free methods is based on smooth particle hydrodynamic procedures [1,2]. These techniques, also called free lagrangian methods, are typically used for problems involving large motions of solids and moving free surfaces in fluids. A second class of mesh free methods derive from generalized finite difference (GFD) techniques [3,4]. Here the approximation around each point is typically defined in terms of Taylor series expansions and the discrete equations are found by using point collocation. Among a third class of mesh free techniques we find the so called diffuse element (DE) method [5], the element free Galerking (EFG) method [6,7] and the reproducing kernel particle (RKP) method [8,9]. These three methods use local interpolations for defining the approximate field around a point in terms of values in adjacent points, whereas the discretized system of equations is typically obtained by integrating the Galerkin variational form over a suitable background grid.

The *finite point method* (FPM) proposed in [10–15] is a truly meshless procedure. The approximation around each point is obtained by using standard moving least square techniques similarly as in DE and EFG methods. The discrete system of equations

is obtained by sampling the governing differential equations at each point as in GFD methods.

The basis of the success of the FPM for solid and fluid mechanics applications is the *stabilization* of the discrete differential equations. The stable form found by the *finite calculus* procedure presented in [16–21] corrects the errors introduced by the point collocation procedure, mainly next to the boundary segments. In addition, it introduces the necessary stabilization for treating high convection effects and it also allows equal order velocity-pressure interpolations in fluid flow problems [19,21].

The content of the chapter is structured as follows. In the next section the basis of the FPM approximation is described. The stabilized governing equations for incompressible flows derived using the finite calculus (FIC) approach are then presented. Next a three step semi-implicit fractional solution scheme using the FPM approximation is described in some detail. Finally, examples of the efficiency and accuracy of the stabilized FPM for numerical solution of incompressible flow problems are presented, namely the analysis of a driven cavity flow, the solution of a backwards facing step, the analysis of a submerged cylinder and the aerodynamic study of a NACA airfoil.

2 INTERPOLATION IN THE FPM

Let Ω_i be the interpolation domain (cloud) of a function $u(x)$ and let s_j with $j = 1, 2, \cdots, n$ be a collection of n points with coordinates $x_j \in \Omega_i$. The unknown function u may be approximated within Ω_i by

$$u(x) \cong \hat{u}(x) = \sum_{l=1}^{m} p_l(x)\alpha_l = \mathbf{p}(x)^T \boldsymbol{\alpha} \tag{1}$$

where $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \cdots \alpha_m]^T$ and vector $\mathbf{p}(x)$ contains typically monomials, hereafter termed "base interpolating functions", in the space coordinates ensuring that the basis is complete. For a 2D problem we can specify

$$\mathbf{p} = [1, x, y]^T \quad \text{for} \quad m = 3 \tag{2}$$

and

$$\mathbf{p} = [1, x, y, x^2, xy, y^2]^T \quad \text{for} \quad m = 6 \qquad \text{etc.} \tag{3}$$

Function $u(x)$ can now be sampled at the n points belonging to Ω_i giving

$$\mathbf{u}^h = \left\{ \begin{array}{c} u_1^h \\ u_2^h \\ \vdots \\ u_n^h \end{array} \right\} \cong \left\{ \begin{array}{c} \hat{u}_1 \\ \hat{u}_2 \\ \vdots \\ \hat{u}_n \end{array} \right\} = \left\{ \begin{array}{c} \mathbf{p}_1^T \\ \mathbf{p}_2^T \\ \vdots \\ \mathbf{p}_n^T \end{array} \right\} \boldsymbol{\alpha} = \mathbf{C}\boldsymbol{\alpha} \tag{4}$$

where $u_j^h = u(x_j)$ are the unknown but sought for values of function u at point j, $\hat{u}_j = \hat{u}(x_j)$ are the approximate values, and $\mathbf{p}_j = \mathbf{p}(x_j)$.

In the FE approximation the number of points is chosen so that $m = n$. In this case $\mathbf{C}$ is a square matrix. The procedure leads to the standard shape functions in the FEM [22].

If $n > m$, $\mathbf{C}$ is no longer a square matrix and the approximation can not fit all the u_j^h values. This problem can be simply overcome by determining the $\hat{u}$ values by minimizing the sum of the square distances of the error at each point weighted with a function $\varphi(x)$ as

$$J = \sum_{j=1}^{n} \varphi(x_j)(u_j^h - \hat{u}(x_j))^2 = \sum_{j=1}^{n} \varphi(x_j)(u_j^h - \mathbf{p}_j^T \boldsymbol{\alpha})^2 \tag{5}$$

with respect to the $\boldsymbol{\alpha}$ parameters. Note that for $\varphi(x) = 1$ the standard least square (LSQ) method is reproduced.

Function $\varphi(x)$ is usually built in such a way that it takes a unit value in the vecinity of the point i typically called "star node" where the function (or its derivatives) are to be computed and vanishes outside a region Ω_i surrounding the point. The region Ω_i can be used to define the number of sampling points n in the interpolation region. A typical choice for $\varphi(x)$ is the normalized Gaussian function and this has been chosen in the examples shown in the paper. Of course $n \geq m$ is always required in the sampling region and if equality occurs no effect of weighting is present and the interpolation is the same as in the LSQ scheme.

Standard minimization of eq.(5) with respect to $\boldsymbol{\alpha}$ gives

$$\boldsymbol{\alpha} = \bar{\mathbf{C}}^{-1}\mathbf{u}^h \qquad , \qquad \bar{\mathbf{C}}^{-1} = \mathbf{A}^{-1}\mathbf{B} \tag{6}$$

$$\begin{aligned} \mathbf{A} &= \sum_{j=1}^{n} \varphi(x_j)\mathbf{p}(x_j)\mathbf{p}^T(x_j) \\ \mathbf{B} &= [\varphi(x_1)\mathbf{p}(x_1), \varphi(x_2)\mathbf{p}(x_2), \cdot\varphi(x_n)\mathbf{p}(x_n)] \end{aligned} \tag{7}$$

The final approximation is obtained by substituting $\boldsymbol{\alpha}$ from eq.(6) into (1) giving

$$\hat{u}(x) = \mathbf{p}^T\bar{\mathbf{C}}^{-1}\mathbf{u}^h = \mathbf{N}^T\mathbf{u}^h = \sum_{j=1}^{n} N_j^i u_j^h \tag{8}$$

where the "shape functions" for the i-th star node are

$$N_j^i(x) = \sum_{l=1}^{m} p_l(x)\bar{C}_{lj}^{-1} = \mathbf{p}^T(x)\bar{\mathbf{C}}^{-1} \tag{9}$$

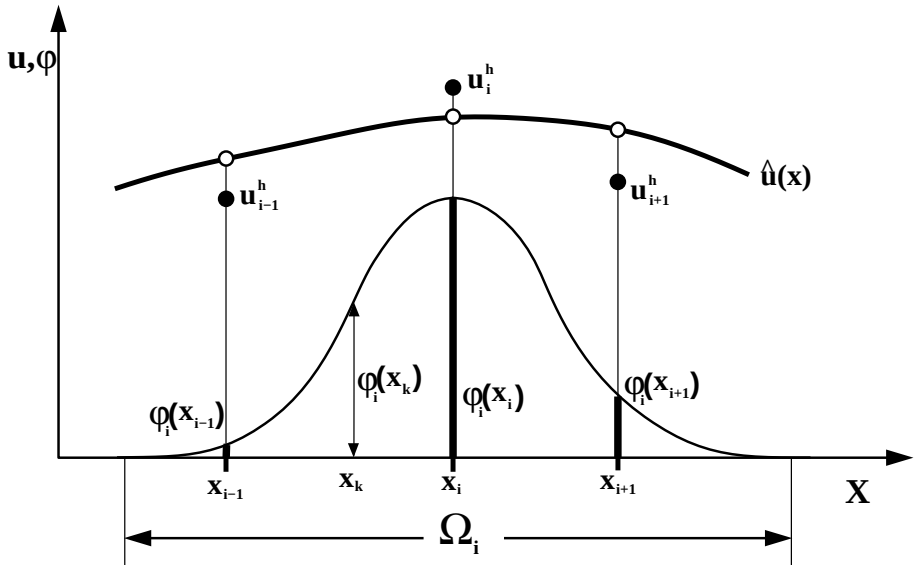


Figure 1. Fixed weighting least square procedure

It must be noted that accordingly to the least square character of the approximation

$$u(x_j) \simeq \hat{u}(x_j) \neq u_j^h \tag{10}$$

i.e. the local values of the approximating function do not fit the nodal unknown values. Indeed $\hat{u}$ is the true approximation for which we shall seek the satisfaction of the differential equation and the boundary conditions and u_j^h are simply the unknown parameters sought.

The weighted least square approximation described above depends on a great extend on the shape and the way to apply the weighting function. The simplest way is to define a fixed function $\varphi(x)$ for each of the Ω_i interpolation domains [11,12].

Let $\varphi_i(x)$ be a weighting functions satisfying (Figure 1)

$$\begin{aligned} \varphi_i(x_i) &= 1 \\ \varphi_i(x) &\neq 0 \qquad x \in \Omega_i \\ \varphi_i(x) &= 0 \qquad x \notin \Omega_i \end{aligned} \tag{11}$$

Then the minimization square distance becomes

$$J_i = \sum_{j=1}^{n} \varphi_i(x_j)(u_j^h - \hat{u}(x_j))^2 \quad \text{minimum} \tag{12}$$

The expression of matrices **A** and **B** coincide with eq.(7) with $\varphi(x_j) = \varphi_i(x_j)$.

Note that according to (1), the approximate function $\hat{u}(x)$ is defined in each interpolation domain Ω_i. In fact, different interpolation domains can yield different shape functions N_j^i. As a consequence a point belonging to two or more overlapping interpolation domains has different values of the shape functions which means that $N_j^i \neq N_j^k$. The interpolation is now multivalued within Ω_i and, therefore for any useful approximation a decision must be taken limiting the choice to a single value. Indeed, the approximate function $\hat{u}(x)$ will be typically used to provide the value of the unknown function $u(x)$ and its derivatives in only specific regions within each interpolation domain. For instance by using point collocation we may limit the validity of the interpolation to a single point x_i. It is precisely in this context where we have found this meshless method to be more useful for practical purposes [10–15].

3 STABILIZED FPM USING A FINITE CALCULUS APPROACH

Finite element solution of the incompressible Navier-Stokes equations with the classical Galerkin method may suffer from numerical instabilities from two main sources. The first is due to the advective-diffusive character of the equations which induces oscillations for high values of the velocity. The second source has to do with the mixed character of the equations which limits the choice of finite element interpolations for the velocity and pressure fields.

Solutions of these two problems have been extensively sought in the last years. Compatible velocity-pressure interpolations satisfying the inf-sup condition emanating from the second problem above mentioned have been used. In addition, the advective operator has been modified to include some "upwinding" effects [22–30]. Recent procedures based on Galerkin Least Square [31,32], Characteristic Galerkin [33,34], Variational Multiscale [35–37] and Residual Free Bubbles [38–40] techniques allow equal order interpolation for velocity and pressure by introducing a Laplacian of pressure term in the mass balance equation, while preserving the upwinding stabilization of the momentum equations. Most of these methods lack enough stability in the presence of sharp layers transversal to the velocity. This defficiency is usually corrected by adding new "shock capturing" stabilization terms to the already stabilized equations [41–43]. The computation of the stabilization parameters in all these methods is based in "ad hoc" generalizations of the parameters for the 1D linear advective-diffusive-reactive problem [44,45].

This paper presents a different point view for deriving stabilized a finite point method for incompressible flow problems. The starting point are the stabilized form of the governing differential equations derived via a *finite calculus* (FIC) procedure. This technique first presented in [16,17] is based on writting the different balance equations over a domain of finite size and retaining higher order terms. These terms incorporate the ingredients for the necessary stabilization of any transient and steady state numerical solution *already at the differential equations level*. Application of the MLS interpolation

and point collocation to the consistently modified differential equations for the fluid flow problem leads to a stabilized system of discretized equations which overcomes *the two problems* above mentioned, i.e. the advective type instability and that due to lack of compatibility between the velocity and pressure fields.

For the sake of preciseness the basic ideas of the FIC method are given next.

3.1 Basic concept of the finite increment calculus (FIC) method

Let us consider a sourceless transient problem over a one dimensional domain AB of length L (Figure 2). The balance of flux q over a domain of finite size belonging to L can be written as

$$q_A - q_B = 0 \tag{13}$$

where A and B are the end points of the finite size domain of length h. As usual q_A and q_B represent the values of the flux q at points A and B, respectively.

For instance, in an 1D advective-diffusive problem the flux $q = -cu\phi + k\frac{d\phi}{dx}$, where ϕ is the transported variable (i.e. the temperature in a thermal problem), u is the advective velocity and c and k are the advective and diffusive material parameters, respectively.

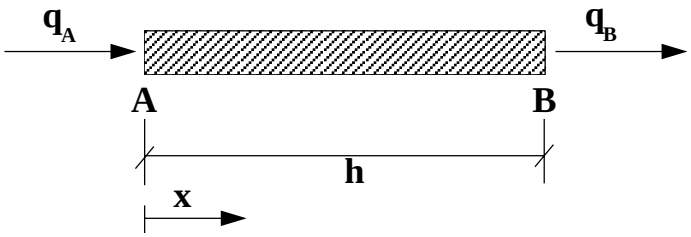


Figure 2. Equilibrium of fluxes in a finite balance domain

The flux q_A can be expressed in terms of the values at point B by the following Taylor series expansion

$$q_A = q_B - h\frac{\partial q}{\partial x}|_B + \frac{h^2}{2}\frac{d^2q}{dx^2}|_B + Oh^3 \tag{14}$$

Substituting (14) into (13) gives after simplification and neglecting cubic terms in h

$$\frac{dq}{dx} - \underline{\frac{h}{2}\frac{dq}{dx}} = 0 \tag{15}$$

where all terms are evaluated at the arbitrary point B.

Eq. (15) is the *finite* form of the balance equation over the domain AB. The underlined term in eq.(15) introduces the necessary stabilization for the discrete solution

of eq.(15) using *any* numerical technique. Distance h is the characteristic length of the discrete problem and its value depends on the parameters of discretization method chosen (such as the grid size). Note that for $h \to 0$ the standard infinitesimal form of the balance equation $\left(\frac{dq}{dx} = 0\right)$ is recovered.

Above process can be extended to derive the stabilized balance differencial equations for any problem in mechanics as

$$r_d - \underline{\frac{h_j}{2}\frac{\partial r_i}{\partial x_j}} = 0 \tag{16}$$

where r_i is the standard form of the ith differential equation for the infinitesimal problem, h_j are the dimensions of the domain where balance of fluxes, forces, etc. is enforced, and $j = 1, 2, 3$ for 3D problems. Details of the derivation of eq.(16) for steady-state and transient advective-diffusive and fluid flow problems can be found in [16]. Applications of the FIC approach to the Galerkin finite element solution of these problems are reported in [16–21].

The underlined stabilization terms in eqs.(15) and (16) are a consequence of accepting that the infinitesimal form of the balance equations is an unreachable limit within the framework of a discrete numerical solution. Indeed eqs.(3) or (4) are *not longer valid* for obtaining an analytical solution following traditional integration methods from infinitesimal calculus theory. The meaning of the new stabilized equations makes only sense in the context of a discrete numerical method yielding approximate values of the solution at a finite set of points within the analysis domain. Convergence to the *exact* analytical value at the points will occur only for the limit case of zero grid size (except for some simple 1D problems [16]) which also implies naturally a zero value of the characteristic length parameters.

The FIC formulation presented below for incompressible flows can be considered an extension of that recently developed in [21] for finite element analysis of incompressible Navier-Stokes flows. The set of stabilized governing equations is first discretized in time using a semi-implicit fractional step procedure and then solved in space using the FPM. The stabilized formulation allows the use of an equal order interpolation for the velocities and pressure variables.

3.2 FIC formulation of viscous flow equations

We consider the motion around a body of a viscous incompressible fluid.

The stabilized FIC form of the governing differential equations for the three dimensional (3D) problem can be written as

Momentum

$$r_{m_i} - \underline{\frac{1}{2}h_{mj}\frac{\partial r_{m_i}}{\partial x_j}} - \underline{\frac{1}{2}\delta\frac{\partial r_{m_i}}{\partial t}} = 0 \qquad \text{on } \Omega \quad i, j = 1, 2, 3 \tag{17}$$

Mass balance

$$r_d + \underline{\frac{1}{2} h_{dj} \frac{\partial r_d}{\partial x_j}} = 0 \qquad \text{on } \Omega \quad j = 1, 2, 3 \tag{18}$$

where

$$r_{m_i} = \rho \left[\frac{\partial u_i}{\partial t} + \frac{\partial}{\partial x_j}(u_i u_j) \right] + \frac{\partial p}{\partial x_i} - \frac{\partial \tau_{ij}}{\partial x_j} - b_i \tag{19}$$

$$r_d = \frac{\partial u_i}{\partial x_i} \qquad i = 1, 2, 3 \tag{20}$$

In above u_i is the velocity along the i-th global reference axis, ρ is the (constant) density of the fluid, p is the pressure, b_i are the body forces acting in the fluid and τ_{ij} are the viscous stresses related to the viscosity μ by the standard expression

$$\tau_{ij} = \mu \left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} - \delta_{ij} \frac{2}{3} \frac{\partial u_k}{\partial x_k} \right) \tag{21}$$

The boundary conditions for the stabilized problem are written as

$$n_j \tau_{ij} + t_i + \underline{\frac{1}{2} h_{mj} n_j r_{m_i}} = 0 \qquad \text{on } \Gamma_t \tag{22}$$

$$u_j - u_j^p = 0 \qquad \text{on } \Gamma_u \tag{23}$$

where n_j are the components of the unit normal vector to the boundary and t_i and u_j^p are prescribed tractions and displacements on the boundaries Γ_t and Γ_u, respectively.

The underlined terms in eqs.(17)–(22) introduce the necessary stabilization for the approximated numerical solution.

The *characteristic length* distances h_{mj} and h_{dj} represent the dimensions of the finite domain where balance of momentum and mass. The signs before the stabilization terms in eqs.(17), (19) and (22) ensure a positive value of the characteristic length distances. The parameter δ in eq.(17) has dimensions of time. Details of the derivation of eqs. (17)–(23) can be found in [16,19,21].

Eqs.(17–23) are the starting point for deriving a variety of stabilized numerical methods for solving the incompressible Navier-Stokes equations. It can be shown that a number of standard stabilized finite element methods allowing equal order interpolations for the velocity and pressure fields can be recovered from the modified form of the momentum and mass balance equations given above [16,19].

Alternative form of the mass balance equation

Taking the first derivative of eq.(21) gives (assuming the viscosity μ to be constant)

$$\frac{\partial \tau_{ij}}{\partial x_j} = \mu \Delta u_i + \frac{\mu}{3} \frac{\partial r_d}{\partial x_i} \tag{24}$$

where $\Delta = \frac{\partial^2}{\partial x_i \partial x_i}$ is the Laplacian operator. Substituting eq.(24) into (17) gives after small algebra

$$\frac{\partial r_d}{\partial x_i} = \left(\frac{\mu}{3} + \frac{\rho u_i h_{m_i}}{2} \right)^{-1} \left[\bar{r}_{m_i} - \frac{h_{m_k}}{2} \frac{\partial r_{m_i}}{\partial x_k} + \frac{\rho u_i h_{m_i}}{2} \frac{\partial r_d}{\partial x_i} - \frac{\delta}{2} \frac{\partial r_{m_i}}{\partial t} \right] \quad \text{no sum in } i \tag{25}$$

where

$$\bar{r}_{m_i} = r_{m_i} + \frac{\mu}{3} \frac{\partial r_d}{\partial x_i} \tag{26}$$

and r_{m_i} is given by eq.(19).

Inserting eq.(25) into eq.(18) gives

$$r_d + c_i \left(\bar{r}_{m_i} - \frac{h_{m_k}}{2} \frac{\partial r_{m_i}}{\partial x_k} + \frac{\rho u_i h_{m_i}}{2} \frac{\partial r_d}{\partial x_i} - \frac{\delta}{2} \frac{\partial r_{m_i}}{\partial t} \right) = 0 \quad \text{no sum in } i \tag{27}$$

with

$$c_i = \left(\frac{2\mu}{3 h_{d_i}} + \frac{\rho u_i h_{m_i}}{h_{d_i}} \right)^{-1} \qquad \text{no sum in } i \tag{28}$$

Eq.(27) can be rewritten as

$$r_d - g_{ii} \frac{\partial^2 p}{\partial x_i \partial x_i} + r_p = 0 \tag{29}$$

where

$$r_p = c_i \bar{r}_{m_i} - g_{ij} \frac{\partial}{\partial x_j} \left(r_{m_i} - \delta_{ij} \frac{\partial p}{\partial x_i} \right) + \frac{\rho u_i h_{m_i}}{2} \frac{\partial r_d}{\partial x_i} - \frac{\delta}{2} \frac{\partial r_{m_i}}{\partial t} \quad \text{no sum in } i \tag{30}$$

and

$$g_{ij} = \left(\frac{4\mu}{3 h_{d_i} h_{m_j}} + \frac{2\rho u_i h_{m_i}}{h_{d_i} h_{m_j}} \right)^{-1} \quad \text{no sum in } i \tag{31}$$

Note that for $h_{m_i} = h_{m_j} = h$ where h is a typical grid dimension (i.e. the average size of a cloud of points), the value of g_{ii} is simply

$$g_{ii} = \left(\frac{4\mu}{3h^2} + \frac{2\rho u_i}{h} \right)^{-1} \tag{32}$$

The stabilization parameter g_{ii} has now the form traditionally used in the Galerkin Least Square formulation for the viscous (Stokes) limit ($u_i = 0$) and the inviscid (Euler) limit ($\mu = 0$) and deduced from ad-hoc extensions of the 1D advective-diffusive problem [25–46]. Note, however, that the general form of the stabilization parameter g_{ii} is deduced here from the general FIC formulation without further extrinsic assumptions.

Indeed, the precise computation of the characteristic length values is crucial for the practical applications of above stabilized expressions. This topic is dealt with on Section 7.

4 FRACTIONAL STEP APPROACH

The momentum equations (17) are first discretized in time using the following scheme

$$u_i^{n+1} = u_i^n - \frac{\Delta t}{\rho}\Big[\rho\frac{\partial (u_i u_j)^n}{\partial x_j} + \frac{\partial p^{n+1}}{\partial x_i} - \frac{\partial \tau_{ij}^n}{\partial x_j} - b_i^n - \frac{h_{m_k}^n}{2}\frac{\partial r_{m_i}^n}{\partial x_k} - \frac{\delta^n}{2}\frac{\partial r_{m_i}^n}{\partial t}\Big] \tag{33}$$

Eq.(33) is now split into the two following equations

$$u_i^* = u_i^n - \frac{\Delta t}{\rho}\Big[\rho\frac{\partial (u_i u_j)}{\partial x_j} - \frac{\partial \tau_{ij}}{\partial x_j} - b_i - \frac{h_{m_k}}{2}\frac{\partial r_{m_i}}{\partial x_k} - \frac{\delta}{2}\frac{\partial r_{m_i}}{\partial t}\Big]^n \tag{34}$$

$$u_i^{n+1} = u_i^* - \frac{\Delta t}{\rho}\frac{\partial p^{n+1}}{\partial x_i} \tag{35}$$

Note that the sum of eqs.(34) and (35) gives the original form of eq.(33).

Substituting eq.(35) into the stabilized mass balance equation (29) gives the standard Laplacian of pressure form

$$\left(\frac{\Delta t}{\rho} + g_{ii}^n\right)\frac{\partial^2 p^{n+1}}{\partial x_i \partial x_i} = r_d^* + r_p^n \tag{36a}$$

where

$$r_d^* = \frac{\partial u_i^*}{\partial x_i} \tag{36b}$$

Standard fractional step procedures neglect the contribution from the terms involving g_{ii} in eq. (36a). These terms have an additional stabilization effect which improves the numerical solution when the values of Δt are small. Note that for $\Delta t \to 0$ the term g_{ii} introduces the necessary stability in the laplacian equation, thereby overcoming the Babuska-Brezzi conditions and allowing for equal order interpolation of the velocities and pressure variables [22].

A typical solution in time includes the following steps.

Step 1. Solve explicitely for the so called fractional velocities u_i^* using eq. (33).

Step 2. Solve for the pressure field p^{n+1} solving the laplacian equation (36a).

Step 3. Compute the velocity field u_i^{n+1} for each mesh node using eq.(35)

5 NUMERICAL SOLUTION USING THE FPM

The implementation of the three step scheme described in previous section in the context of the FPM is straight forward. Eq. (8) is used to define the approximation of velocities and pressures within each cloud of point Ω_i as

$$\hat{u}_m = \sum_{j=1}^{n} N_j^i u_{mj}^h; \quad m = 1, 2, 3 \quad \text{for 3D} \tag{39}$$

$$\hat{p} = \sum_{j=1}^{n} N_j^i p_j^h \tag{40}$$

where $(\hat{\cdot})$ denotes approximate values and the shape functions N_j^i were defined in eq.(9).

Direct substitution of eqs.(39) and (40) into the stabilized governing equations described in previous section gives the following numerical scheme for computation of the point parameters $u_{m_j}^h$ and p_j^h.

Step 1. Computation of fractional velocities

Compute *explicitely* the fractional velocities at each point k in the domain as

$$(\hat{u}_i^*)_k = (\hat{f}_i^n)_k; \quad k = 1, \ldots, N; \quad i = 1, 2, 3 \tag{41}$$

in which N is the total number of points in the domain and

$$(\hat{f}_i^n)_k = \left\{ \hat{u}_i^n - \frac{\Delta t}{\rho} \left[\rho \frac{\partial (\hat{u}_i \hat{u}_j)}{\partial x_j} - \frac{\partial \hat{\tau}_{ij}}{\partial x_j} - b_i - \frac{h_{m_j}}{2} \frac{\partial \hat{r}_{m_i}}{\partial x_j} - \frac{\delta}{2} \frac{\partial \hat{r}_{m_i}}{\partial t} \right]^n \right\}_k \tag{42}$$

where $(\hat{\cdot})$ denotes approximate values.

Once the values of $\hat{u}_i^*$ have been obtained, the parameters $u_{m_j}^h$ can be computed at each point by solving the following system of equations

$$(\hat{u}_m^*)_k = \sum_{j=1}^{n} N_j^k u_{m_j}^h, \quad k = 1, \ldots, N \tag{43}$$

Eq.(43) is a system of N equations with N unknowns from where the parameters $u_{m_j}^h, \quad j = 1, \ldots, N$ can be found. These parameters are needed to compute the

derivatives of the velocity field in steps 2 and 3. Indeed the solution of eq.(43) must be repeated for every component of the velocity vector (i.e. $m = 1, 2, 3$ for 3D problems).

Step 2. Computation of pressures at time $n+1$

Compute the pressure field at time $n+1$ by solving eq.(36a). Substituting eqs. (40) and (43) into (36a) and sampling this equation at each point in the domain gives

$$\mathbf{K}^n(\mathbf{p}^h)^{n+1} = \hat{\mathbf{r}}_d^* + \hat{\mathbf{r}}_p^n \tag{44}$$

where (for 2D problems)

$$K_{kj}^n = (\frac{\Delta t}{\rho} + \hat{g}_{ii}^n)\left(\frac{\partial^2 N_j^k}{\partial x_1^2} + \frac{\partial^2 N_j^k}{\partial x_2^2}\right) \tag{45}$$

$$\begin{aligned} \hat{r}_{d_k}^* =& \hat{c}_i \hat{\bar{r}}_{m_i} - \hat{g}_{ij}\frac{\partial}{\partial x_j}\left(\hat{r}_{m_i} - \delta_{ij}\frac{\partial \hat{p}}{\partial x_i}\right) + \frac{\rho \hat{u}_i h_{m_i}}{2}\frac{\partial \hat{r}_d}{\partial x_i} - \frac{\delta}{2}\frac{\partial \hat{r}_{m_i}}{\partial t} \quad \text{no sum in } i \\ \hat{r}_{p_k}^* =& \left[\frac{\partial \hat{u}_i^*}{\partial x_i}\right] \end{aligned} \tag{46}$$

Eq.(46) provides a system of equations from which the pressure parameters $(p_k^h)^{n+1}$ can be found at each point k.

Step 3. Computation of velocities at time $n+1$

The final step is the explicit computation of the velocities in each point at time $n+1$. Substituting the known values of $\hat{u}_i$ and $\hat{p}^{n+1}$ at each point into eq.(35) gives

$$(\hat{u}_i^{n+1}) = \left[\hat{u}_i^* - \frac{\Delta t}{\rho}\frac{\partial \hat{p}^{n+1}}{\partial x_i}\right]_k; \quad k = 1, \ldots, N \tag{47}$$

Note that the derivatives of the approximate functions $\hat{u}_i$ and $\hat{p}$ are computed by direct differentiation of the expressions (39) and (40), i.e.

$$\begin{aligned} \frac{\partial \hat{u}_m}{\partial x_l} &= \sum_{j=1}^{n} \frac{\partial N_j^i}{\partial x_l} u_{m_j}^h \\ \frac{\partial \hat{p}}{\partial x_l} &= \sum_{j=1}^{n} \frac{\partial N_j^i}{\partial x_l} p_j^h \end{aligned} \tag{48}$$

The steps 1–3 described above are repeated for every new time increment.

A local time step size for each point in the domain can be used to speed up the search of the steady state solution. The local time step is defined as $\Delta t_i = \frac{d_i}{2|\mathbf{u}_i|}$, where d_i is the minimum distance from a star point to any of its neighbourghs in the cloud. Note however that the full transient solution requires invariably the use of a global time step Δt_g equal for all nodes and defined as $\Delta t_g = \min(\Delta t_i), \quad i = 1, \ldots, N$.

6 BOUNDARY CONDITIONS

Prescribed tractions on the Neumann boundary Γ_t, (eq.(22)) or prescribed velocities at the Dirichlet boundary Γ_u (eq.(23)) may be imposed.

During the fractional step solution, the first explicit step is solved without imposing any boundary conditions. During the second step, two kinds of boundary conditions may be imposed: on boundaries where the normal velocity is imposed to the value u_n^p, eq.(23) reads using (35)

$$u_n^p = u_i^* n_i - \frac{\Delta t}{\rho} \frac{\partial p^{n+1}}{\partial x_i} n_i \tag{49}$$

Eq.(49) is a Neumann boundary condition for the pressure equation (36a). This equation is imposed in the FPM during the pressure computation (step 2) as a new equation for all points k belonging to the part of the boundary Γ_u where the normal velocity is prescribed.

On outflow boundaries with $n_j \sigma_{ij} = 0$ the pressure is imposed to a constant value, i.e. $p = 0$. In the FPM, essential boundary conditions such as $p = 0$ are imposed using the definition of the function itself via eq.(40) as

$$\hat{p}_i = \sum_{j=1}^{n} N_j^i p_j^h = 0 \tag{50}$$

Equation (50) is sampled at the points located at a boundary where $p = 0$.

During the third step the velocities are computed at all points using eq.(47) at all points within the analysis domain. In points where a velocity is imposed as an essential boundary condition, the imposed velocity value is asigned directly to the point. Next, the nodal velocity parameters $u_{m_j}^h$ are computed by solving the same system of equations described by eq.(43). For points over Neumann boundaries, in particular on boundaries where the tractions are prescribed to zero, the discretized form of eq.(22), i.e.

$$n_j \hat{\tau}_{ij} + \frac{1}{2} h_{m_j} n_j \hat{r}_{m_i} = 0 \tag{51}$$

is used for computing the velocities at the boundary points.

7 COMPUTATION OF THE STABILIZATION PARAMETERS

Accurate evaluation of the stabilization parameters is one of the crucial issues in stabilized methods. Most of existing methods use expressions which are direct extensions of the values obtained for the simplest 1D case. It is also usual to accept the so called "streamline upwind" assumption. It can be shown that this is equivalent to admit that vector $\mathbf{h}_m$ has the direction of the velocity field [16,19]. This unnecessary restriction leads to instabilities when sharp layers transversal to the velocity direction are present. This additional defficiency is usually corrected by adding a shock capturing or crosswind stabilization term [41–43]. In the FIC approach the crosswind stabilization is naturally introduced into the discretized equations through the general form of the characteristic length vector.

Let us first assume for simplicity that the stabilization parameters for the mass balance equations are the same than those for the momentum equations. This implies

$$\mathbf{h}_m = \mathbf{h}_d = \mathbf{h} \tag{52}$$

The problem remains now finding the value of the characteristic length vectors $\mathbf{h}$. Indeed, the components of $\mathbf{h}$ introduce the necessary stabilization along the streamline and transversal directions to the flow.

Excellent results have been obtained in all examples by using the same value of the characteristic length vector for each momentum equation defined by

$$\mathbf{h} = h_s \frac{\mathbf{u}}{|\mathbf{u}|} + h_c \frac{\boldsymbol{\nabla} u}{|\boldsymbol{\nabla} u|} \tag{53}$$

where $u = |\mathbf{u}|$ and h_s and h_c are the "streamline" and "cross wind" length parameters given by

$$h_s = \max(\mathbf{l}_j^T \mathbf{u})/|\mathbf{u}| \tag{54}$$

$$h_c = \max(\mathbf{l}_j^T \boldsymbol{\nabla} u)/|\boldsymbol{\nabla} u| \quad , \quad j = 1, 2, \cdots n \tag{55}$$

where $\mathbf{l}_j$ are the vectors linking each node in the cloud with the star node.

Note that the cross-wind terms in eq.(53) account for the effect of the gradient of the velocity field in the stabilization parameters. This is an standard assumption in most "shock-capturing" stabilization procedures [41–43].

Regarding the time stabilization parameter δ and in eq.(17) the value $\delta = \Delta t$ has been taken for the solution of the examples presented in the paper.

8 NUMERICAL EXAMPLES

The following examples have been solved with the FPM presented in previous section using a Gaussian weighting function in the WLS approximation and quadratic interpolation ($m = 6$) for the both the velocities and the pressure. Typically each cloud contains nine points ($n = 9$) which are chosen using a quadrant search scheme (i.e. the star node plus the two closest points within each quadrant are selected) [11-13].

8.1 Driven cavity flow at $Re = 1000$

This is a classical test problem to evaluate the behaviour of any fluid dynamic algorithm. A viscous flow is confined in a square cavity while one of its edges slides tangentially. The boundary conditions are $u = v = 0$ in 3 edges and $u = 1$, $v = 0$ on the upper edge. The problem is solved with the FPM using the distribution of 3,329 points shown in Figure 3. Initially, except at the edge, the velocity is set to zero everywhere including at the nodes located at the left and right top corners (ramp condition).

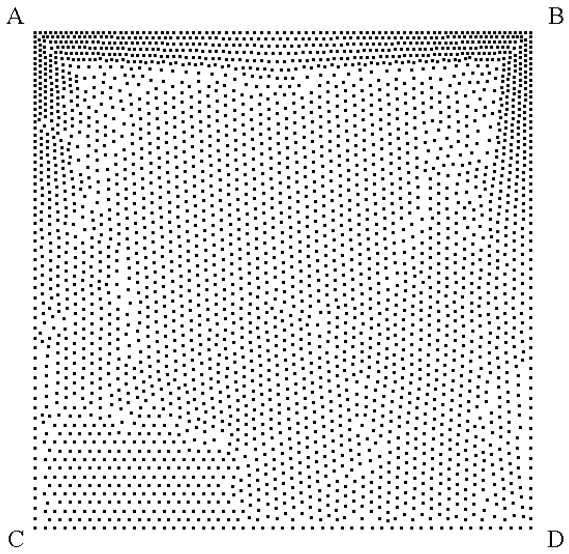


Figure 3. Driven cavity flow. Distribution of 3,329 points. Boundary conditions $\mathbf{u} = 0$ at edges AC, CD and BD and points A and B. $u = 1$ and $v = 0$ over the interior of line AB

Numerical results are shown in Figures 4, 5 and 6 for $Re = 1000$. Figures 4 and 5 show the velocity and pressure contours, respectively. The FPM results are compared with experimental results obtained by Ghia *et al.* [46] showing the velocity x computed along a vertical central cut (Figure 6). The comparison is satisfactory.

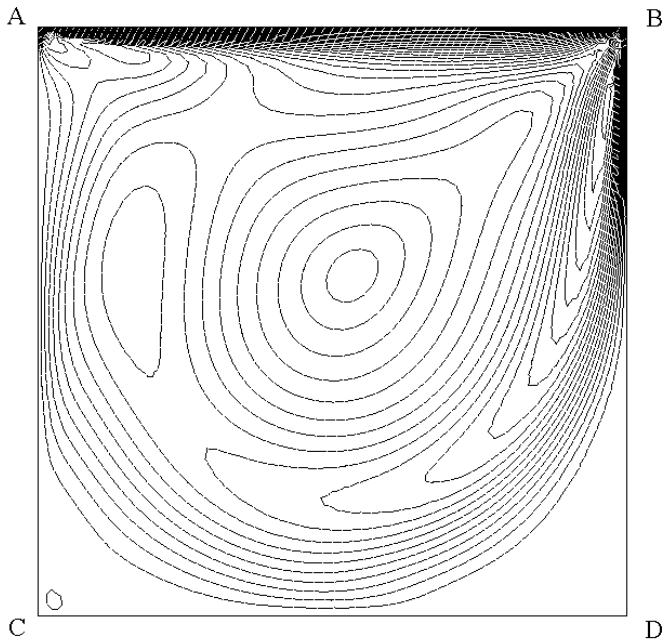


Figure 4. Driven cavity flow. Velocity contours for $Re = 1000$

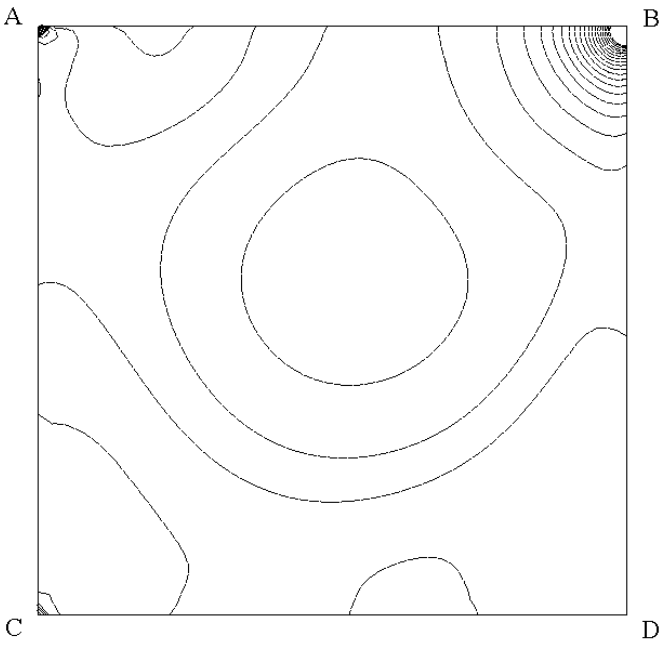


Figure 5. Driven cavity flow. Pressure contours for $Re = 1000$

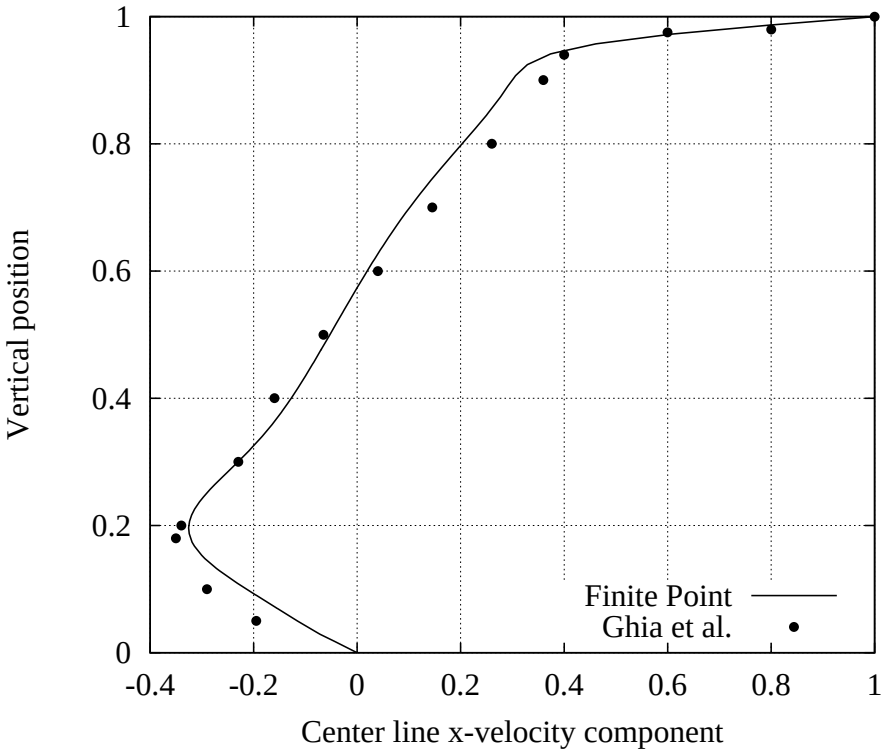


Figure 6. Driven cavity flow. Horizontal velocity distribution over the center line

8.2 Backwards facing step at $Re = 389$

In this example, the flow is contrained to move in a 2D domain which presents a backwards step. The domain dimensions are presented in Figure 7. The step is one half the width of the inflow.

At the inflow a constant velocity profile is fixed while at the outflow the pressure is prescribed, being the velocity free. The non-slip condition is used at the walls, except for the two inflow points, where the constant inflow velocity is imposed. No volume forces are present.

The distribution of 8,462 points used near the step is represented on Figure 8. In the rest of the domain a regular distribution of points is used.

Once the stationary state is reached, the solution shows horizontal velocities represented on Figures 9 and 10 for two planes located at $x = 2.55$ S and $x = 6.11$ S from the step. The FPM results are compared with experimental results presented on ref.[47] showing an excellent agreement.

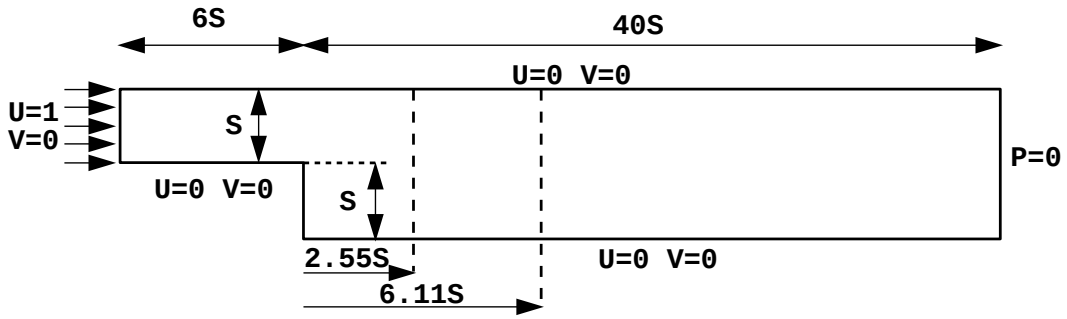


Figure 7. Backwards facing step. Geometry and boundary conditions

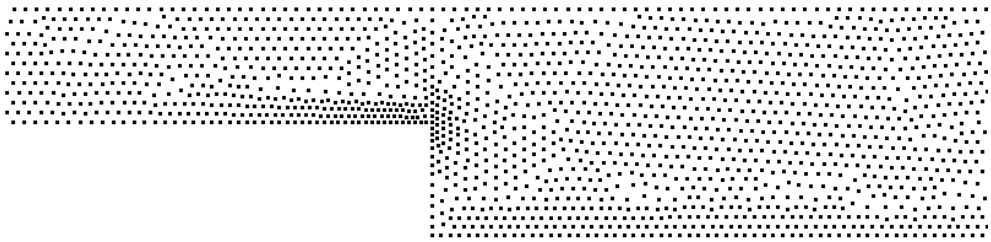


Figure 8. Backwards facing step. Distribution of 8,462 points

8.3 2D viscous flow around a cylinder

Figure 11 shows the geometry of the analysis domain and the boundary conditions. The problem was solved for $Re = 100$ assuming laminar flow conditions. An arbitrary grid of 9418 points was chosen for the analysis (Figure 12). The transient analysis was run for 10000 time steps. The steady state solution was found after 18000 time steps. Note that a full period in the solution requires just 321 time steps.

Figure 13 shows the velocity contour lines at four different times. Note the oscilatory character of the solution. The time evolution of the lift force is shown in Figure 14. The oscillation period deduced from the computation is 6.01 sec. This value compares well with the experimental result of 5.98 sec. ($\simeq 0.5\%$ error) reported by Roshko [48].

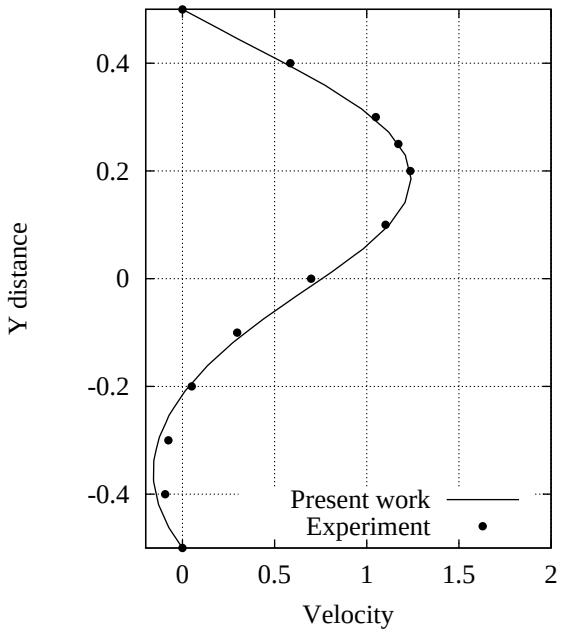


Figure 9. Backwards facing step. Horizontal velocity distribution along a vertical line at $x = 2.55$ S

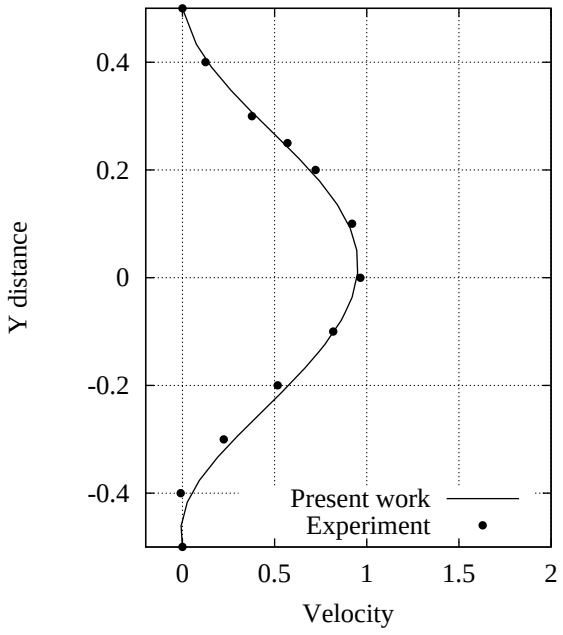


Figure 10. Backwards facing step. Horizontal velocity along a vertical line at $x = 6.11$ S

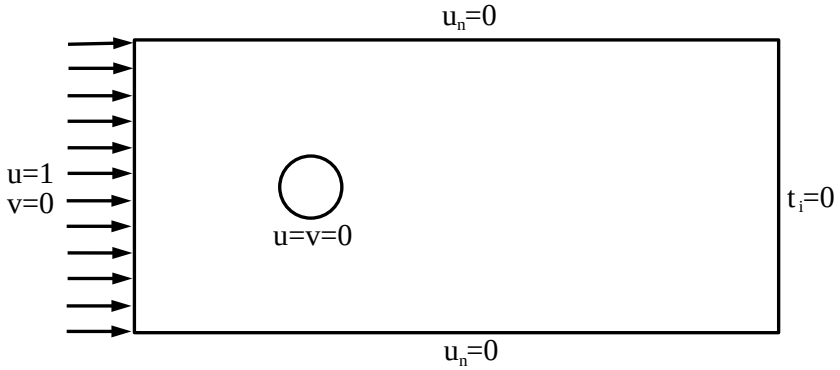


Figure 11. 2D flow around a cylinder. Analysis domain and boundary conditions. $Re = 100$. Boundary tractions (t_i) are assumed to be zero at the exit boundary

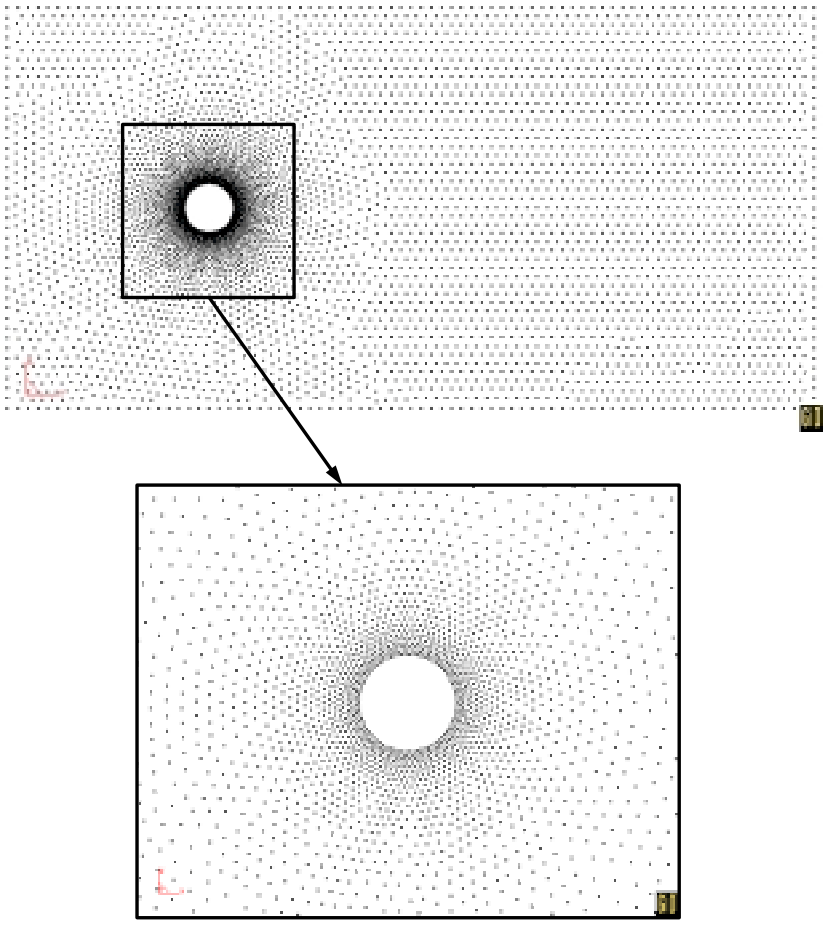


Figure 12. Grid of 9418 points used for analysis of the 2D flow around a cylinder

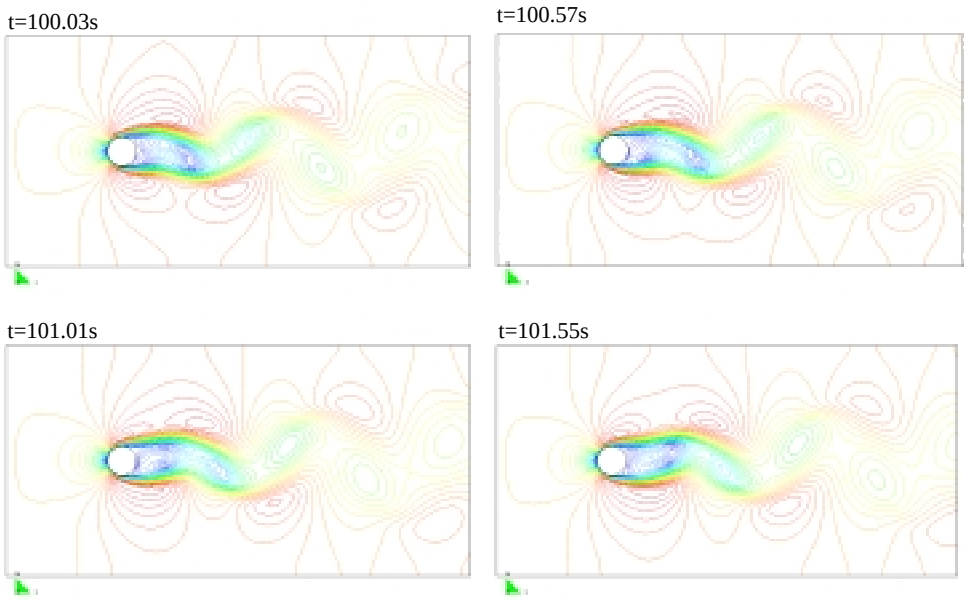


Figure 13. 2D flow around a cylinder ($Re = 100$). Velocity streamlines at different times

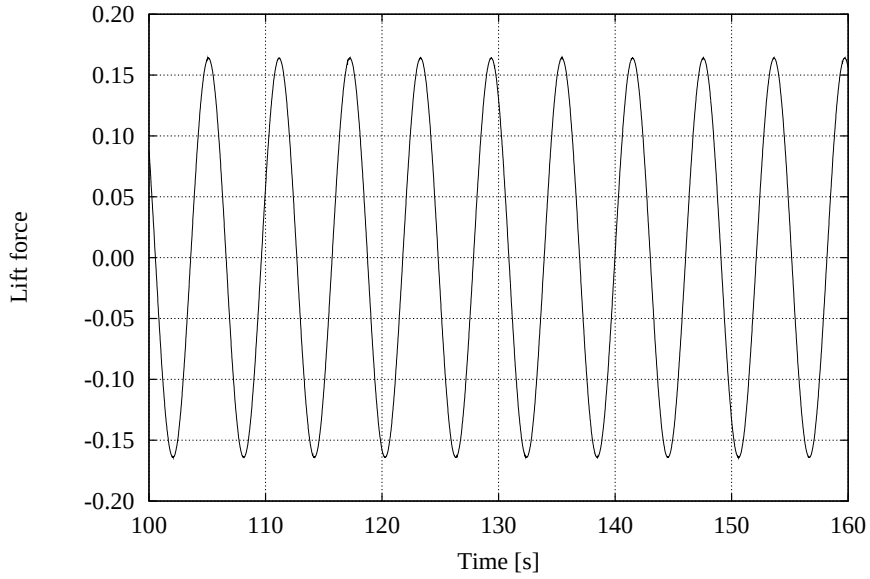


Figure 14. 2D flow around a cylinder. Time evolution of lift force

8.4 2D viscous flow around a Naca airfoil

The viscous flow around a NACA 0012 airfoil for an angle of attack of zero degrees and $Re = 10000$ was analyzed. Laminar flow conditions were again assumed.

Figure 15 shows the geometry of the domain and the boundary conditions. The grid of 14249 points chosen is shown on Figure 16. A finer layer of 972 points was used around the airfoil to capture viscous effects as shown in the figure.

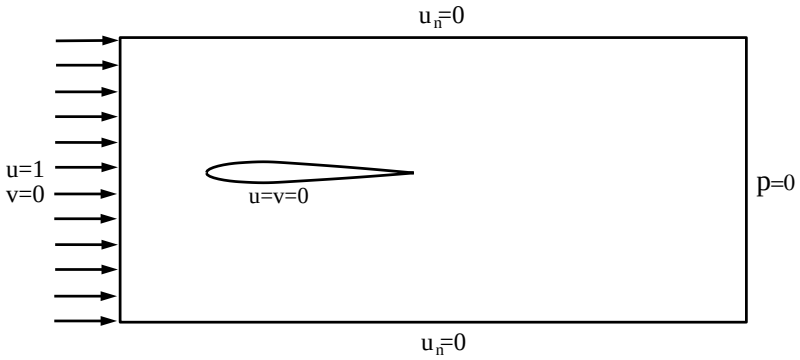


Figure 15. 2D flow around a NACA airfoil. $\alpha = 0°$, $Re = 10000$. Analysis domain and boundary conditions

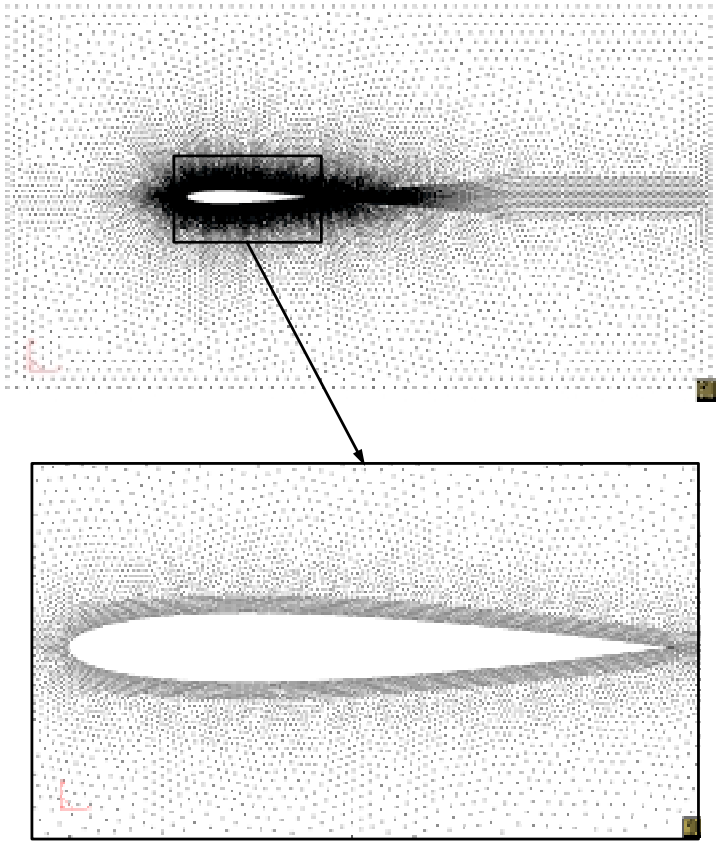


Figure 16. Distribution of 14249 points for analysis of a NACA airfoil. Detail of boundary layer of 972 point to capture viscous effects

Figure 17 shows some numerical results of the velocity streamlines for the steady state situation. Note the well developed wake at the back of the airfoil. A close up of the streamlines next to the airfoil showing the boundary layer developed is also presented.

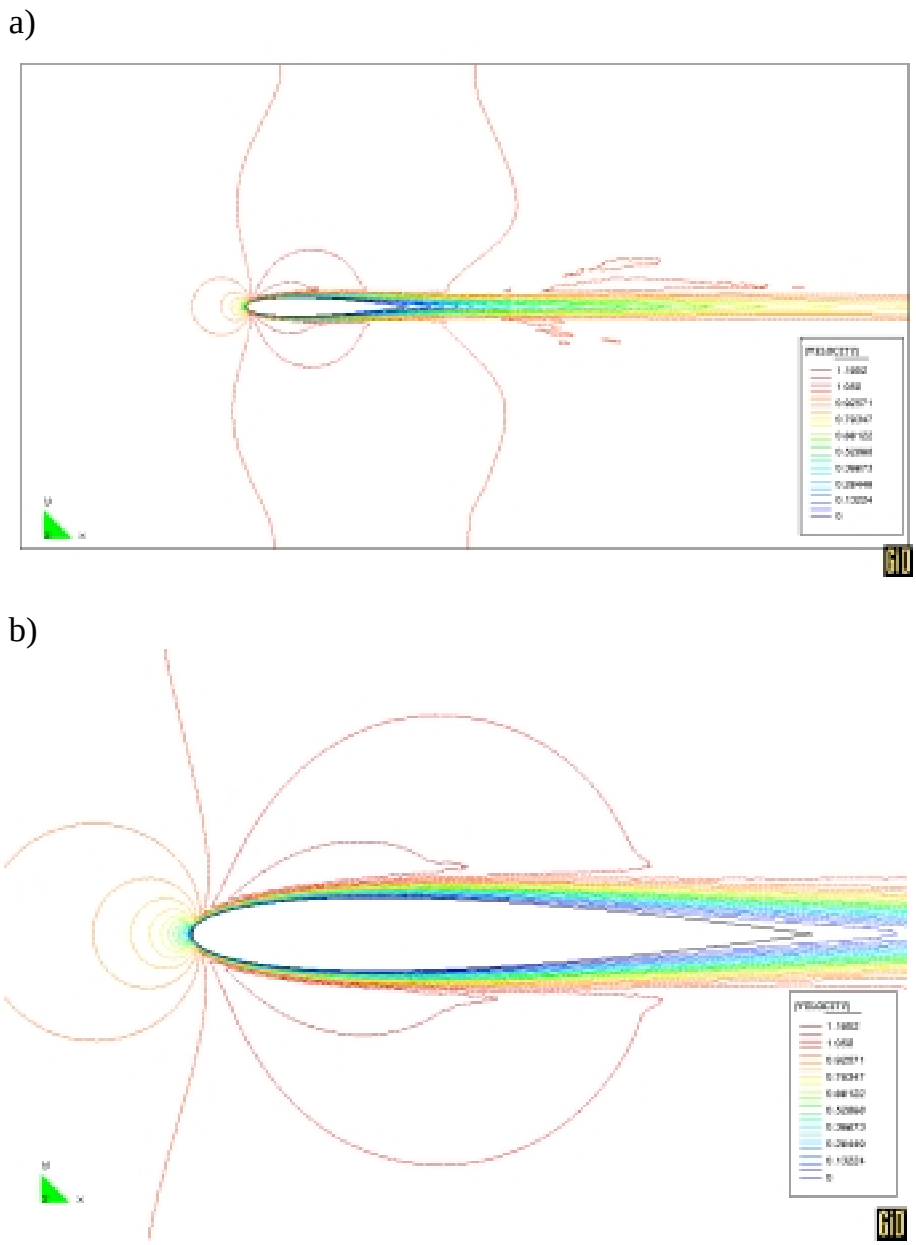


Figure 17. 2D analysis of a NACA airfoil. Velocity streamlines at steady state for $\alpha = 0°$ and $Re = 10000$

9 FINAL CONCLUSSIONS

The stabilized equations for a viscous incompressible fluid using the finite calculus procedure are the basis for deriving a stabilized finite point method for the meshless solution of incompressible flows. The three step semi-implicit fractional scheme provides a simple and accurate procedure for both transient and steady state solutions using equal order interpolation for the velocities and the pressure. The stabilized FPM is a promising technique for the practical meshless solution of industrial flow problems.

REFERENCES

[1] J.J. Monaghan, "Smoothed particle hydrodynamics: Some recent improvement and applications", *Annu. Rev. Astron. Physics*, **30**, p. 543, (1992).

[2] P.W. Randles and L.D. Libersky, "Smoothed particle hydrodynamics: Some recent improvement and applications", *Appl. Mech. Engng.*, **139**, p. 175, (1996).

[3] N. Perrone and R. Kao, *A general finite difference method for arbitrary meshes.*, *Comp. Struct*, **5**, pp. 45-47, (1975).

[4] T. Liszka and J. Orkisz, "The finite difference method at arbitrary irregular grids and its application in applied mechanics", *Comp. Struct.*, **11**, pp. 83-95, (1980).

[5] B. Nayroles, G. Touzot and P. Villon, "Generalizing the FEM: Diffuse approximation and diffuse elements. Comput. Mechanics", **10**, pp. 307-318, (1992).

[6] T. Belytschko, Y. Lu and L. Gu, "Element free Galerkin methods.", *Int. J. Num. Meth. Engng.*, **37**, pp. 229-56, (1994).

[7] J. Dolbow and T. Belytschko, "An introduction to programming the meshless element free Galerkin method", *Archives of Comput. Meth. in Engng.*, **5**, (3), pp. 207–241, (1998).

[8] W.K. Liu, S. Jun, S. LI, J. Adee and T. Belytschko, "Reproducing Kernel particle methods for structural dynamics", *Int. J. Num. Meth. Engng.*, **38**, pp. 1655–1679, (1995).

[9] W.K. Liu, Y Chen, S. Jun, J.S. Chen, T. Belytschko, C. Pan, R.A. Uras and C.T. Chang, "Overview and applications of the Reproducing Kernel particle method", *Archives of Comput. Meth. in Engng.*, Vol. **3**, (1), pp. 3–80, (1996).

[10] E. Oñate, S. Idelsohn, O.C. Zienkiewicz and T. Fisher, "A finite point method for analysis of fluid flow problems", *Proceedings of the 9th Int. Conference on Finite Element Methods in Fluids*, Venize, Italy, pp. 15-21, October, (1995).

[11] E. Oñate, S. Idelsohn, O.C. Zienkiewicz and R.L. Taylor, "A finite point method in computational mechanics. Applications to convective transport and fluid flow", *Int. J. Num. Meth. Engng.*, Vol. **39**, pp. 3839–3866, (1996).

[12] E. Oñate, S. Idelsohn, O.C. Zienkiewicz and R.L. Taylor, "A stabilized finite point method for analysis of fluid mechanics's problems", *Comput. Meth. in Appl. Engng.*, Vol. **139**, pp. 1-4, pp. 315–347, (1996).

[13] E. Oñate and S. Idelsohn, "A mesh free finite point method for advective-diffusive transport and fluid flow problems", *Computational Mechanics*, **21**, pp. 283–292, (1998).

[14] E. Oñate, C. Sacco and S. Idelsohn, "A finite point method for incompressible flow problems", *Computing and Visualization in Sciences*, **3**, 67–75, 2000.

[15] R. Löhner, C. Sacco, E. Oñate and S. Idelsohn, "A finite point method for compressible flow", to be published in *Int. J. Num. Meth. Engng.*, 2000.

[16] E. Oñate, "Derivation of stabilized equations for advective-diffusive transport and fluid flow problems", *Comput. Meth. Appl. Mech. Engng.*, Vol. 151, 1-2, pp. 233–267, (1998).

[17] E. Oñate, J. García and S. Idelsohn, "Computation of the stabilization parameter for the finite element solution of advective-diffusive problems", *Int. J. Num. Meth. Fluids*, Vol. 25, pp. 1385–1407, (1997).

[18] E. Oñate, J. García and S. Idelsohn, "An alpha-adaptive approach for stabilized finite element solution of advective-diffusive problems with sharp gradients", *New Adv. in Adaptive Comp. Met. in Mech.*, P. Ladeveze and J.T. Oden (Eds.), Elsevier, (1998).

[19] E. Oñate, "A finite element method for incompressible viscous flows using a finite increment calculus formulation", *Comput. Meth. Appl. Mech. Engng.*

[20] E. Oñate and M. Manzán, "A general procedure for deriving stabilized space-time finite elements for advective-diffussive problems", *Int. J. Num. Meth. in Fluids.*

[21] E. Oñate and J. García, "A finite element method for fluid-structure interaction with surface waves using a finite calculus formulation", *Comp. Meth. Appl. Mech. Engng.*, **182**, 355–370, 2000.

[22] O.C. Zienkiewicz and R.L. Taylor, "*The finite element method.*", 5th Edition, Arnold, (2000).

[23] A. Brooks and T.J.R. Hughes, "Streamline upwind/Petrov-Galerkin formulation for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations", *Comput. Meth. Appl. Mech. Engng*, **32**, 199–259, 1982.

[24] T.J.R. Hughes and M. Mallet, "A new finite element formulations for computational fluid dynamics: III. The generalized streamline operator for multidimensional advective-diffusive systems", *Comp. Meth. Appl. Mech. Engng.*, **58**, pp. 305–328, 1986.

[25] P. Hansbo and a. Szepessy, "A velocity-pressure streamline diffusion finite element method for the incompressible Navier-Stokes equations", *Comp. Meth. Appl. Mech. Engng.*, **84**, 175–192, 1990.

[26] T.J.R. Hughes, L.P. Franca and M. Balestra, "A new finite element formulation for computational fluid dynamics. V Circumventing the Babuska-Brezzi condition: A stable Petrov-Galerkin formulation of the Stokes problem accomodating equal order interpolations", *Comp. Meth. Appl. Mech. Engng.*, **59**, 85–89, 1986.

[27] L.P. Franca and S.L. Frey, "Stabilized finite element methods: II. The incompressible Navier-Stokes equations", *Comput. Meth. Appl. Mech. Engn*, Vol. **99**, pp. 209–233,

1992.

[28] T.J.R. Hughes, G. Hauke and K. Jansen, "Stabilized finite element methods in fluids: Inspirations, origins, status and recent developments", in: *Recent Developments in Finite Element Analysis*. A Book Dedicated to Robert L. Taylor, T.J.R. Hughes, E. Oñate and O.C. Zienkiewicz (Eds.), (International Center for Numerical Methods in Engineering, Barcelona, Spain, pp. 272–292, 1994.

[29] M.A. Cruchaga and E. Oñate, "A finite element formulation for incompressible flow problems using a generalized streamline operator", *Computer Methods in Applied Mechanics and Engineering*, 143, 49–67, 1997.

[30] M.A. Cruchaga and E. Oñate, "A generalized streamline finite element approach for the analysis of incompressible flow problems including moving surfaces", *Computer Methods in Applied Mechanics and Engineering*, 173, 241–255, 1999.

[31] T.J.R. Hughes, L.P. Franca and G.M. Hulbert, "A new finite element formulation for computational fluid dynamics: VIII. The Galerkin/least-squares method for advective-diffusive equations", *Comput. Meth. Appl. Mech. Engng.*, **73**, pp. 173–189, 1989.

[32] T.E. Tezduyar, S. Mittal, S.E. Ray and R. Shih, "Incompressible flow computations with stabilized bilinear and linear equal order interpolation velocity–pressure elements", *Comp. Meth. Appl. Mech. Engng.*, **95**, 221–242, 1992.

[33] O.C. Zienkiewicz and R. Codina, "A general algorithm for compressible and incompressible flow. Part I: The split characteristic based scheme", *Int. J. Num. Meth. in Fluids*, **20**, 869-85, (1995).

[34] O.C. Zienkiewicz, K. Morgan, B.V.K. Satya Sai, R. Codina and M. Vázquez, "A general algorithm for compressible and incompressible flow. Part II: Tests on the explicit form", *Int. J. Num. Meth. in Fluids*, **20**, No. 8-9, 886-913, 1995.

[35] T.J.R. Hughes, "Multiscale phenomena: Greens functions, subgrid scale models, bubbles and the origins of stabilized methods", *Comput. Meth. Appl. Mech. Engng*, Vol. **127**, pp. 387–401, 1995.

[36] R. Codina, "A stabilized finite element method for generalized stationary incompressible flows", Publication PI-148, CIMNE, Barcelona, February 1999.

[37] R. Codina and J. Blasco, "Stabilized finite element method for the transient Navier-Stokes equations based on a pressure gradient operator". To appear in *Computer Methods in Appl. Mech. Engng.*

[38] F. Brezzi, M.O. Bristeau, L.P. Franca, M. Mallet and G. Rogé, "A relationship between stabilized finite element methods and the Galerkin method with bubble functions", *Comput. Meth. Appl. Mech. Engn.*, Vol. **96**, pp. 117–129, 1992.

[39] F. Brezzi, D. Marini and A. Russo, "Pseudo residual-free bubbles and stabilized methods", *Computational Methods in Applied Sciences '96*, J. Periaux *et. al.* (Eds.), J. Wiley, 1996.

[40] F. Brezzi, L.P. Franca, T.J.R. Hughes and A. Russo, "$b = \int g$", *Comput. Meth. Appl. Mech. Engn.*, **145**, 329–339, 1997.

[41] T.J.R. Hughes and M. Mallet, "A new finite element formulations for computational fluid dynamics: IV. A discontinuity capturing operator for multidimensional advective-diffusive system, *Comput. Meth. Appl. Mech. Engng.*, **58**, 329–336, 1986.

[42] A.C. Galeão and E.G. Dutra do Carmo, "A consistent approximate upwind Petrov-Galerkin method for convection dominated problems", *Comput. Meth. Appl. Mech. Engng.*, **68**, 83–95, 1988.

[43] R. Codina, "A discontinuity-capturing crosswind dissipation for the finite element solution of the convection-diffusion equation", *Comput. Meth. Appl. Mech. Engng.*, **110**, 325–342, 1993.

[44] R. Codina, "Comparison of some finite element methods for solving the diffusion-convection-reaction equation". *Comp. Meth. Appl. Mech. Engng.*, **188**, 61–82, 2000.

[45] R. Codina, "On stabilized finite element methods for linear systems of convection-diffusion-reaction equation", Publication CIMNE PI-162, December 1997. To appear in *Computer Meth. Appl. Mech. Engng.*

[46] U. Ghia, K. Ghia and C. Shin, "High-*Re* solutions for incompressible flow using the Navier-Stokes equation and a multi-grid method", *J. Comp. Phys.*, Vol. **48**, pp. 387–441, (1982).

[47] B.F. Armaly, F. Durst, J.C.F. Pereira and B. Schönung, "Experimental and theoretical investigations of backword-facing step flow", *J. Fluid Mech.*, Vol. **127**, pp. 473–496, (1983).

[48] A. Roshko, "On the development of turbulent wakes form vortex streets", NACA Technical Report No. 1191, National Advisory Committee for Aeronautics, USA, (1954).

Shells and finite elements: from classicism to modernism

Juhani Pitkäranta

Institute of Mathematics
Helsinki University of Technology
P.O. Box 1100, FIN-02015 HUT, Finland

1 Introduction

The problem of mathematically or numerically modelling shell deformations is a rather unique problem of engineering science: Equations similar to the linear *shell equations* that describe the deformation of a shell as an elastic body are hardly met anywhere else. When modelling shell deformations by finite elements, special difficulties also arise that are characteristic to shell problems only.

Shells are anyway very common structures in both nature and engineering. The mathematical theory of shells also got its first inspiration from an old engineering strucure, the church bell. The pioneering work, as motivated by this problem in particular, was done by Love in 1888 [7]. Since then, the shell theory has expanded to what may now be called the *classical shell theory*. This is documented in books written by numerous authors including Love [8], Vlasov [18], Novozhilov [10] and many others.

The final words in the classical linear shell theory were set by Koiter [4] and Naghdi [9] in the early 1960's. Around the same time, the computer modelling of shell deformations using *finite elements* took its first steps. This new 'modernism', however, was not the child of the 'classicism' of shell theory. Rather, the finite element designers went back to the basic equations of linear elasticity, known since the early 1800's, and designed the finite element models on the basis of those equations directly. The classical shell theory, as developed after the late 1800's, was thus more or less ignored – and is still ignored – in the finite element engineering.

The early finite element models of shells, however, did not work as desired. Unexpected numerical phenomena appeared in the context of finite element discretizations, and it has taken a long time to understand the cause of such phenomena. Even today, shell problems are still the most challenging problems of structural mechanics for the finite element designer, and the results of numerical models cannot always be trusted.

The failure of finite element engineering calls for mathematical error analysis. The finite element theory has been developed intensively since the 1970's, but the theory is

mostly based on far simpler problems than the shell problem. Only fairly recently there have been serious attempts to extend the theory to cover the more specific problems met in shell modelling. In this context, as it turns out, the results of classical shell theory are needed once again. First of all, the understanding of the basic mathematical nature of shell deformations is necessary for finite element error analysis, and this understanding is still based primarily on the classical shell theory. Secondly, the numerical phenomena arising in finite element models are understood more easily when the finite element approximation is thought of in the context of the simpler classical shell models.

Thus, although the classical shell theory is still of little use in the finite element modelling practice, a renessaince of the classical theory is seen in the finite element theory. Our aim in this review is to demonstrate the new interaction of the 'classicism' and the 'modernism' in shell modelling. We start from the shell problem formulated as a 3D elastic problem and discuss first the dimension reductions of classical shell theory and their connection to the finite element approximations. We define then an extremely simplified version of the classical shell models to be called the *mathematical shell model*. Here all the unnecessary details of the classical shell theories are left out, while still preserving the main characteristics of the original shell problem. We proceed to outline the leading characteristic features of shell deformations on the basis of the simplified model. Based on this information we further outline the leading steps required in the finite element error analysis and derive some general error bounds.

2 The shell problem

The starting point of our study is the classical problem of linear elasticity where an elastic body, consisting of homegeneous isotropic material and occupying a region $\Omega \subset R^3$, is deformed under a given load and kinematic constraints. According to the energy principle, the deformation is obtained by minimizing the total energy

$$\mathcal{F}(\mathbf{u}) = \frac{1}{2}\int_\Omega \boldsymbol{\sigma}(\mathbf{u}) : \boldsymbol{\varepsilon}(\mathbf{u})\, d\Omega - \mathcal{L}(\mathbf{u}) \tag{2.1}$$

over the kinematicallly admissible displacement fields $\mathbf{u} = (u_1, u_2, u_3)$ of the body. Here $\mathcal{L}(\mathbf{u})$ is the load functional (potential energy of the load) and the first term is the strain energy (deformation energy), expressed in terms of the stress tensor $\boldsymbol{\sigma}$ and the strain tensor $\boldsymbol{\varepsilon}$. In case of a homogeneous isotropic material the stress and strain are related by the generalized Hooke law $\boldsymbol{\sigma} = \lambda\, tr\boldsymbol{\varepsilon}\, \mathbf{I} + 2\mu\, \boldsymbol{\varepsilon}$, where $\mathbf{I}$ is the identity tensor and λ, μ are the Lamé parameters of the material, defined in terms of the Young modulus E and Poisson ratio ν $(0 \le \nu < 1/2)$ as

$$\lambda = \frac{E\nu}{(1+\nu)(1-2\nu)}, \quad \mu = \frac{E}{2(1+\nu)}.$$

Taking into account the stress-strain relations, the leading strain energy term in Eq. (2.1) takes the quadratic form

$$\begin{aligned}\mathcal{A}(\mathbf{u},\mathbf{u}) &= \int_\Omega \boldsymbol{\sigma}(\mathbf{u}) : \boldsymbol{\varepsilon}(\mathbf{u})\, d\Omega \\ &= \int_\Omega \left\{ \lambda\, [tr\boldsymbol{\varepsilon}(\mathbf{u})]^2 + 2\mu\, \boldsymbol{\varepsilon}(\mathbf{u}) : \boldsymbol{\varepsilon}(\mathbf{u}) \right\}\, d\Omega. \end{aligned} \tag{2.2}$$

As is well known, the actual mathematical character of the above problem depends strongly on the geometric shape of the body Ω. Indeed, much of the classical linear elasticity consists of developing specific theories for specific engineering objects like beams, bars, rods, shafts, plates, etc. [8] Such objects have in common that their smallest characteristic dimension is much less than the diameter of the body. What we consider here is the most challenging one of the thin-body problems of classical linear elasticity: the problem of a curved thin shell. An elastic body is a shell when it is a smoothly deformed thin plate, so that $\Omega = \boldsymbol{\Phi}(\Omega')$ where $\boldsymbol{\Phi} :\ R^3 \mapsto R^3$ is a smooth map and $\Omega' = \omega \times (-d/2, d/2)$, where $\omega \subset R^2$ is an open set and d is much smaller than the smallest characteristic dimension of ω. (As the latter we may take, say, the diameter of the largest disc fully contained in ω.) More specifically we consider a shell of constant thickness d and assume that the mapping $\boldsymbol{\Phi}$ is of the form

$$\boldsymbol{\Phi}(x,y,z) = \boldsymbol{\Phi}(x,y,0) + z\,\mathbf{n}(x,y), \quad (x,y,z) \in \Omega',$$

where $\mathbf{n}$ is the normal to the shell midsurface Γ defined as

$$\Gamma = \Gamma_0, \quad \Gamma_z = \{\boldsymbol{\Phi}(x,y,z) \mid (x,y) \in \omega\} = \boldsymbol{\Phi}\,(\,\omega \times \{z\}).$$

The coordinates x, y then parametrize the shell midsurface Γ, and x, y, z act as the natural curvilinear coordinate system on Ω. Upon transforming the displacements and strains into this coordinate system, the strain energy integral (2.2) takes the form

$$\int_\Omega \{\,..\,\}\, d\Omega = \int_{-d/2}^{d/2} \left(\int_{\Gamma_z} \{\,..\,\}\, d\Gamma_z \right) dz.$$

For a *shell of revolution*, it is easy to define the coordinates x, y as the *principal curvature* coordinates, so that the coordinate lines are principal curvature lines. Consider the special case of a *cylindrical shell*. Let (r, φ, x) be the cylindrical coordinate system where x is the axial coordinate. In the above notation one has then $r = R + z$, where R is the radius of the shell midsurface. Choosing $y = R\varphi$, we have defined the coordinates x, y on Γ as the principal curvature coordinates associated to a unit metric tensor. The mapping $\boldsymbol{\Phi}$ may then be defined as

$$\begin{aligned}\boldsymbol{\Phi}(x,y,z) &= (x, r\sin(y/R), r\cos(y/R)) \\ &= (x, R\sin(y/R), R\cos(y/R)) + z\,(0, \sin(y/R), \cos(y/R)) \\ &= \boldsymbol{\Phi}(x,y,0) + z\,\mathbf{n}(x,y).\end{aligned}$$

The displacement vector field $\mathbf{u}$ on the cylindrical shell is expressed in the chosen curvilinear coordinate system as

$$\mathbf{u}(x,y,z) = u_1(x,y,z)\,\mathbf{e}_x + u_2(x,y,z)\,\mathbf{e}_y + u_3(x,y,z)\,\mathbf{e}_z,$$

where $\mathbf{e}_x = (1,0,0)$, $\mathbf{e}_z = \mathbf{n}$, and $\mathbf{e}_y = \mathbf{n} \times \mathbf{e}_x$. The components of the symmetric strain tensor $\boldsymbol{\varepsilon}(\mathbf{u})$ are then given by

$$\begin{aligned}
&\varepsilon_{11} = \frac{\partial u_1}{\partial x}\,, \quad \varepsilon_{22} = \frac{\partial u_2}{\partial y} + \frac{u_3}{r}\,, \quad \varepsilon_{33} = \frac{\partial u_3}{\partial z}\,, \\
&\varepsilon_{12} = \varepsilon_{21} = \frac{1}{2}\left(\frac{\partial u_1}{\partial y} + \frac{\partial u_2}{\partial x}\right), \quad \varepsilon_{13} = \varepsilon_{31} = \frac{1}{2}\left(\frac{\partial u_1}{\partial z} + \frac{\partial u_3}{\partial x}\right), \\
&\varepsilon_{23} = \varepsilon_{32} = \frac{1}{2}\left(\frac{\partial u_2}{\partial z} + \frac{\partial u_3}{\partial y} - \frac{u_2}{r}\right).
\end{aligned}$$

3 Dimension reduction: The 2D m-models

The classical engineering and mathematical theory of shells, has produced a wide literature and a large variety of specific *shell models*, where the original 3D elastic problem is reduced to a 2-dimensional one. The current understanding of the mathematical nature of shell deformation is almost enirely based on such 2D models, so let us outline, how the dimension reduction is carried out in shell theory.

All dimension reductions may be viewed mathematically as being based on one primary assumption: Supposer that the displacement field $\mathbf{u}$ varies as a function of x, y (i.e., along the midsurface) in some *characteristic length scale* L. Then it is assumed that

$$t_{\text{eff}} = \frac{d}{L} \ll 1. \tag{3.1}$$

Parameter t_{eff} is the *effective thickness* (dimensionless thickness) of the shell. In practice, shell deformations are typically of multiscale nature, so that many length scales of different orders of magnitude are present simultaneously. Asuumption (3.1) is then related to the *scale resolution* ability of the dimension reduction model: The model is expected to resolve a given length scale the more accurately the less the corresponding effective thickness t_{eff}. The most typical length scales that may arise from the geometric or physical setup of the problem are

- $L = D = diam(\Omega)$, or $L = a =$ characteristic dimension of ω
- $L = R =$ smallest principal radius of curvature on Γ
- $L =$ length scale of load variation
- $L = \lambda =$ wavelength of an eigenmode (vibration, buckling)

A characteristic length scale may also arise from the Fourier mode analysis of the deformation state, in which case

- L = characteristic wavelength of a Fourier mode.

An important and rather unique feature of shell deformations is the presence of *boundary layers* that can have extremely wide range. Boundary layers in shell deformations are rather complex multiscale phenomena that arise from the inherent asymptotic nature of the linear elastic equation on a thin curved body. Under simplifying assumptions, it is possible to perform a Fourier expansion of the layer into exponentially deacying modes [17]. For each mode, the natural length scale is set as

- L = length scale of defay (Fourier layer mode).

Again the dimension reduction model is expected to approximate well layer modes decaying relatively slowly so that Eq. (3.1) holds. There are also layer modes with $L \sim d$ thus violating the assumption. Such components of the 3D elastic deformation cannot be captured by 2D shell models.

Under the basic assumption (3.1) there are two systematic ways of deriving approximate 2D shell models from the 3D elastic model of a shell: Either one simplifies the differential equatioins of 3D elasticity in the assumed coordinates using methods of asymptotic analysis, or one proceeds from the energy principle (or from variational formulations) and expands the displacement field as a polynomial in the normal coordinate z to the shell midsurface. The former approach, although rarely systematic, was the dominant approach in the classical theory of shells, written mostly before the era of computers, cf. [8, 18, 10]. Here we choose the more 'modern' energy approach, as this is parallel with finite element modelling ideas.

We will call the dimension reduction model an *m-model* when based on expanding the displacement field $\mathbf{u} = (u_1, u_2, u_3)$ in the above curvilinear coordinate system (x, y, z) as

$$\left\{\begin{array}{rcl} u_1(x,y,z) &=& u(x,y) + \displaystyle\sum_{i=1}^{m} z^i\, \theta_i(x,y), \\ u_2(x,y,z) &=& v(x,y) + \displaystyle\sum_{i=1}^{m} z^i\, \psi_i(x,y), \\ u_3(x,y,z) &=& w(x,y) + \displaystyle\sum_{i=1}^{m+1} z^i\, \xi_i(x,y), \end{array}\right. \tag{3.2}$$

an minimizing the 3D energy (2.2) with this Ansatz. Here the stopping indices of the expansions are chosen so that each displacement component produces a contribution up to the oder $\mathcal{O}(|z|^m)$ to the diagonal strains ε_{ii} and hence also to the diagonal stresses σ_{ii}. In the resulting model the integrals over z can be evaluated, either exactly or approximately, so one ends up in a 2D shell model where $u, v, w, \theta_i, \psi_i, \xi_i$ constitute the $3m+4$ components of the generalized displacement field, now defined over the shell midsurface Γ.

In the numerical modelling of shells, the above m-model resembles closely a finite element model where the shell is discretized using a single layer of solid finite elements of degree $m+1$ in z. In such a model, the curvilinear coordinates x, y, z acts as the natural reference coordinate system for defining the reference (master) elements on Ω'. In the engineering software, shells are usually modelled in this way, so that no specific 'shell theories' are needed. The transformation of the displacements and strains to the curvilinear coordinate system is neither needed in such models: The element stiffness matrices can be evaluated simply by expressing the displacements and strains in a rectangular coordinate system, then transforming the energy integrals into the reference coordinates, and finally evaluating the transformed integrals numerically.

4 The reduced 3D model

In the above m-model hierarchy, the 7-field 1-model is of special interest, as this is the lowest-order model that captures correctly the asymptotics of shell deformations at the limit of zero thickness. All the the lowest-order 2D shell models found in the classical shell theory may be understood as variations or simplifications of the 1-model. We will also rely on such simplified models when resolving the asymptotics and boundary layers of shell deformations, so let us discuss these simplifications.

The first step to simplify the 1-model is to eliminate the field components ξ_1 and ξ_2 so as to obtain a 5-field model. The elimination is carried out by minimizing the energy first with respect to ξ_1, ξ_2 and eliminating ξ_1 and ξ_2 from the corresponding Euler equations. Under assumption (3.1) the minimization can be carried out approximately by dropping the off-diagonal strina terms, as these are of order $\mathcal{O}(t_{\text{eff}})$ compared with the diagonal terms. To sufficient accuracy one may also approximate $d\Gamma_z$ by $d\Gamma$ when evaluating the volume integrals. After these simplifications, the two Euler equations reduce to the following integral constraints for the normal stress σ_{33} along the midsurface Γ:

$$\int_{-d/2}^{d/2} \sigma_{33}(x,y,z)\,dz = \int_{-d/2}^{d/2} z\,\sigma_{33}(x,y,z)\,dz = 0, \quad (x,y) \in \Gamma. \tag{4.1}$$

These constraints are close to the simple constraint

$$\sigma_{33} = 0, \tag{4.2}$$

which is the famous *plane stress* assumption in classical plate/membrane or shell theories.

Proceeding from Eqs. (4.1), the five-field model is now obtained by solving these equations for ξ_1, ξ_2 and inserting the solutions in the 3D energy expression. Neglecting again small terms of order $\mathcal{O}(t_{\text{eff}})$, the resulting simplified strain energy expression can be rewritten in the original form (2.2) of the 3D energy with the following two modifications enforced:

(i) Set the strain $\varepsilon_{33}(\mathbf{u})$ to zero.

(ii) Redefine the Lamé parameter λ as $\lambda = \dfrac{E\nu}{1-\nu^2}$.

As is well known, the adjustement (ii) of λ is consistent with the plane stress assumption (4.2). Thus we conclude that when obeying the energy formulation of the classical linear elastic problem, it is the linear and quadratic terms in the expansion of the normal displacement component u_3 in (3.2) that effectively enforce the plane stress condition (4.2) on a thin shell.

Once the above modifications within the 3D energy formulation are enforced, the resulting model may be considered a variant of the 3D models as stated, i.e., with no expansion of the displacement field assumed a priori. This is a shell model in itself, often referred to as the *reduced 3D model*. In the numerical modelling of shells, the reduced 3D model is the standard choice in combination with lowest-order element approximations where the expansion of the displacement field is linear in z. A linear expansion is sufficient within the reduced 3D model, since the modifications made in the model approximately take into account the quadratic term in the 1-model.

Below we consider the reduced 3D model primarily as a simplification of the 1-model, and we look for further simplifications of this model in the spirit of classical shell theory. At this point we thus depart from the practice of the finite element modelling the original 3D elastic formulation overrules all specific shell theories. Why classical shell theories are anyhow still important even in the context of finite element models, is basically for mathematical reasons: Simpler 2D shell models are needed to understand the behavior of the 3D finite element algorithms, and in particular to see the possible sources of failure. From this point of view, if not from the programming point of view, the 'classicism' of the shell theory has still something important to offer to a finite element 'modernist'.

5 The classical 2D shell models

We proceed from the reduced 3D model considered as a simplified 1-model where the displacement field is expanded as

$$\left\{\begin{array}{rcl} u_1(x,y,z) &=& u(x,y) - z\,\theta(x,y), \\ u_2(x,y,z) &=& v(x,y) - z\,\psi(x,y), \\ u_3(x,y,z) &=& w(x,y). \end{array}\right. \tag{5.1}$$

Here u, v are the tangential displacements of the midsurface, w is the transverse deflection, and $\theta = -\theta_1$, $\psi = -\psi_1$ are the so called rotations. Assuming this expansion, we can simplify the model further by expanding the geometric coefficients in the strains $\varepsilon_i j$ as a function of z and dropping all terms of order $\mathcal{O}(z^2)$ in the expansions of the strains so obtained. The resulting simplified strains then take the form

$$\begin{aligned} &\varepsilon_{ij}(x,y,z) = \beta_{ij}(x,y) + z\,\kappa_{ij}(x,y), \quad i,j = 1,2, \\ &\varepsilon_{i3}(x,y,z) = \rho_i(x,y), \quad i = 1,2, \\ &\varepsilon_{33}(x,y,z) = 0, \end{aligned}$$

where $\boldsymbol{\beta} = (\beta_{ij})$ and $\boldsymbol{\kappa} = (\kappa_{ij})$ are tensor fields and $\boldsymbol{\rho} = (\rho_i)$ is a vector field defined along the midsurface Γ. These are referred to as the *membrane, bending* and *transverse shear strains*, respectively. These are related linearly to the displacement field components u, v, w, θ, ψ via variable coefficients that depend on the local metric and curvature parameters of Γ. The general expressions of these strains in a tensorial notation were first derived by Naghdi [9]. For our purposes, it will be sufficient to resolve the leading terms of these expressions in special coordinates, which we can do without advanced tensor notation. Consider a point $P \in \Gamma$ and assume coordinates x, y chosen so that $\mathbf{e}_x$ and $\mathbf{e}_y$ are orthogonal at P and that the metric tensor of Γ is the unit tensor at P. Let further

$$b_{11} = \mathbf{e}_x \cdot \frac{\partial \mathbf{n}}{\partial x}\,, \quad b_{22} = \mathbf{e}_y \cdot \frac{\partial \mathbf{n}}{\partial y}\,, \quad b_{12} = b_{21} = \mathbf{e}_x \cdot \frac{\partial \mathbf{n}}{\partial y} = \mathbf{e}_y \cdot \frac{\partial \mathbf{n}}{\partial x}$$

at point P, so that $\mathbf{b} = (b_{ij})$ is the curvature tensor of Γ at P. Then the membrane strains at P can be expanded as

$$\begin{aligned} \beta_{11} &= \frac{\partial u}{\partial x} + b_{11} w + [u, v], \quad , \quad \beta_{22} = \frac{\partial v}{\partial y} + b_{22} w + [u, v], \\ \beta_{12} &= \frac{1}{2}\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right) + b_{12} w + [u, v], \end{aligned} \tag{5.2}$$

the transverse shear strains as

$$\rho_1 = \theta - \frac{\partial w}{\partial x} + [u, v], \quad \rho_2 = \psi - \frac{\partial w}{\partial y} + [u, v], \tag{5.3}$$

and the bending strains as

$$\begin{aligned} \kappa_{11} &= \frac{\partial \theta}{\partial x} + [u, v, w, \theta\, \psi], \quad \kappa_{22} = \frac{\partial \psi}{\partial y} + [u, v, w, \theta, \psi], \\ \kappa_{12} &= \frac{1}{2}\left(\frac{\partial \theta}{\partial y} + \frac{\partial \psi}{\partial x}\right) + [u, v, w, \theta, \psi]. \end{aligned} \tag{5.4}$$

Here we show only the most significant terms that are needed below to characterize the mathematics of shell deformations. The unresolved additional terms denoted by [..] are variable-coefficient linear combinations of the displacement components indicated. In Eq. (5.2) the sign of the curvature tensor is defined so that the principal curvature at $P \in \Gamma$ is positive when the corresponding center of curvature is in the direction $-\mathbf{n}$ from P.

The above assumptions hold at each point P for a cylindrical shell with the principal curvature coordinates (x, y) chosen as above. In this case the the strains are defined

precisely as

$$\begin{aligned}
\beta_{11} &= \frac{\partial u}{\partial x}\,, \quad \beta_{22} = \frac{\partial v}{\partial y} + \frac{w}{R}\,, \quad \beta_{12} = \frac{1}{2}\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right), \\
\rho_1 &= \theta - \frac{\partial w}{\partial x}\,, \quad \rho_2 = \psi - \frac{\partial w}{\partial y} + \frac{v}{R}\,, \\
\kappa_{11} &= \frac{\partial \theta}{\partial x}\,, \quad \kappa_{22} = \frac{\partial \psi}{\partial y} + \frac{1}{R}\,\beta_{22}, \\
\kappa_{12} &= \frac{1}{2}\left(\frac{\partial \theta}{\partial y} + \frac{\partial \psi}{\partial x} + \frac{1}{R}\frac{\partial v}{\partial x}\right) + \frac{1}{R}\beta_{22}.
\end{aligned} \tag{5.5}$$

Using the strain expansion (5), the strain energy of the reduced 3D model may be integrated with respect to z so as to obtain an expression of the form

$$\mathcal{A}(\mathbf{u}, \mathbf{u}) = \int_\Gamma [\,\ldots] \, d\Gamma,$$

where $\mathbf{u}$ now stands for the generalized displacement field (u, v, w, θ, ψ) defined on the midsurface Γ. Dropping the z-dependence of the volume metric by writing $d\Gamma_z = d\Gamma$ causes only an error of order $\mathcal{O}(d/R)$ where R is the least principal radius of curvature on Γ. Assuming this simplification it remains to evaluate the integrals

$$\int_{-d/2}^{d/2} z^i \, dz = \begin{cases} d & \text{for } i = 0, \\ 0 & \text{for } i = 1, \\ d^2/12 & \text{for } i = 2, \end{cases}$$

to conclude that the strain energy can be expressed as

$$\mathcal{A}(\mathbf{u}, \mathbf{u}) = D\,[\,\mathcal{A}_m(\mathbf{u}, \mathbf{u}) + \mathcal{A}_s(\mathbf{u}, \mathbf{u}) + \mathcal{A}_b(\mathbf{u}, \mathbf{u})\,], \tag{5.6}$$

where D is a scaling coefficient defined by

$$D = \frac{Ed}{1 - \nu^2}\,,$$

and $\mathcal{A}_m, \mathcal{A}_s, \mathcal{A}_b$ stand for the scaled membrane, transverse shear and bending energy functionals defined by

$$\begin{aligned}
\mathcal{A}_m(\mathbf{u}, \mathbf{u}) &= \int_\Gamma [\,\nu(\beta_{11} + \beta_{22})^2 + (1 - \nu)(\beta_{11}^2 + 2\beta_{12}^2 + \beta_{22}^2)\,]\, d\Gamma, \\
\mathcal{A}_s(\mathbf{u}, \mathbf{u}) &= \frac{1 - \nu}{2} \int_\Gamma (\rho_1^2 + \rho_2^2)\, d\Gamma, \\
\mathcal{A}_b(\mathbf{u}, \mathbf{u}) &= \frac{d^2}{12} \int_\Gamma [\,\nu(\kappa_{11} + \kappa_{22})^2 + (1 - \nu)(\kappa_{11}^2 + 2\kappa_{12}^2 + \kappa_{22}^2)\,]\, d\Gamma.
\end{aligned} \tag{5.7}$$

The shell model (5.2)–(5.7) is usually referred to as the *Naghdi model*, referring to the systematic derivation of the model in general coordinates x, y by Naghdi [9]). Earlier

incredients of the model are found widely distributed in the literature on plate and shell theories between the 1940's and the 1960's and even in the still earlier theory of beams.

The Naghdi model can be simplified by approximately minimizing the energy with respect to the rotations θ, ψ and eliominating the rotations in this way. The process is very similar to the derivation of the reduced 3D model as outlined above: Under the basic assumption (3.1), the minimizing conditions can be written approximately (omitting terms of order $\mathcal{O}(t_{\text{eff}})$) as

$$\rho_1 = \rho_2 = 0. \tag{5.8}$$

These are the well-known *Kirchhoff-Love constraints*, originally found in plate theory. Upon eliminating the rotations from these constraints as

$$\theta = \frac{\partial w}{\partial x} + [u, v], \quad \psi = \frac{\partial w}{\partial y} + [u, v]$$

results in a simplified energy expression where now $\mathbf{u} = (u, v, w)$ in (5.6), the membrane strains remain unchanged from (5.2), the transverse shear strains vanish according to constraints (5.8), and the bending strains are redefined in terms of u, v, w as

$$\kappa_{11} = \frac{\partial^2 w}{\partial x^2} + [u, v, w], \quad \kappa_{22} = \frac{\partial^2 w}{\partial y^2} + [u, v, w], \quad \kappa_{12} = \frac{\partial^2 w}{\partial x \partial y} + [u, v, w]. \tag{5.9}$$

The model (5.2), (5.8), (5.9), (5.6)–(5.7) is the fundamental 2D shell model, used most often as a basis of further mathematical studies of shell deformations. The model is often named the *Koiter model*, referring to the first systematic derivation in a general coordinate system x, y by Koiter [4, 5]. In principal curvature coordinates, however, the same model was obtained earlier by Vlasov [18], and a number of small variations of the basic model appear in the wide literature on shell theory before the 1960's. The variations (usually presented in princiipal curvature coordinates) typically differ only in how the additional terms $[u, v, w]$ are defined in the expressions (5.9) of the bending strains. Such variations were proposed, e.g., by Novozhilov [10] and, indeed, already by Love in his pioneering work on shell theory in 1888 [7]. For most practical purposes, the variations between the different models are rather irrelevant. We name the Koiter model and its close variants briefly as the *classical 3-field model*. The Naghdi model and its close variants are named analogously as the *classical 5-field model*.

6 The mathematical shell model

The classical 3-field model, the derivation of which was outlined above, is still the basic model of shell theory when the goal is to understand of the mathematical nature of shell deformations. When the goal is to understand the behavior of finite element algorithms, this models goes one step too far, since the Kirchhoff-Love constraints (5.8) is by no means naturally imposed in finite element models. In fact, the difficulty of enforcing constriants of this type is the main source of numerical error amplification in the finite element models

of shell deformations. When analyzing finite element models we thus need to take a step backwards and take the 5-field model as the starting point. The approximations done when deriving this model can be reasonably traced numerically, so that this model is expected to preserve the main characteristics of the original shell problem also from the finite element modelling point of view.

Taking the classical 5-field and 3-field models as the starting point, we perform one final step of simplification in these models. First, we set $\nu = 0$ and scale the deformation energy by setting $D = 1$ in Eq. (5.6). Secondly, we drop the additional terms [..] in the strain expressions (5.2)–(5.4) and (5.9). Third, we assume that the curvature tensor (b_{ij}) is constant on Γ and shorten the notation by writing $a = b_{11}$, $b = b_{22}$, $c = b_{12}$.

It turns out that the above simplifications affect the main characteristic of the shell problem neither mathematically or numerically. The scaling of the energy by setting $D = 1$ is equivalent to scaling the load functional by factor D. The last two assumptions are contradictory, but they hold anyway approximately on a *shallow shell*, the midsurface of which deviates only slightly from a plane. In general, a shell may be considered shallow if

$$\delta = \frac{L}{R} \ll 1,$$

where $L = diam(\Gamma)$ and R is the least principal radius of curvature on Γ. The generic idea in *shallow shell models* is to expand the strains also with respect to δ and keep only the leading terms. The third assumption above is consistent with such an approximation, cf. [17] for a more detailed reasoning. The classical shell theory contains a large number of specific simplified shell models derived by this kind of reasoning.

Under the above simplifying assumptions, the strains of the classical five-field model are written as

$$\begin{aligned}
&\beta_{11} = \frac{\partial u}{\partial x} + aw, \quad \beta_{22} = \frac{\partial v}{\partial y} + bw, \quad \beta_{12} = \frac{1}{2}\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right) + cw, \\
&\rho_1 = \theta - \frac{\partial w}{\partial x}\,, \quad \rho_2 = \psi - \frac{\partial w}{\partial y}\,, \\
&\kappa_{11} = \frac{\partial \theta}{\partial x}\,, \quad \kappa_{22} = \frac{\partial \psi}{\partial y}\,, \quad \kappa_{12} = \frac{1}{2}\left(\frac{\partial \theta}{\partial y} + \frac{\partial \psi}{\partial x}\right).
\end{aligned} \tag{6.1}$$

In the 3-field model the constraints (5.8) are again imposed, so that the bending strains get redefined as

$$\kappa_{11} = \frac{\partial^2 w}{\partial x^2}\,, \quad \kappa_{22} = \frac{\partial^2 w}{\partial y^2}\,, \quad \kappa_{12} = \frac{\partial^2 w}{\partial x \partial y}\,. \tag{6.2}$$

In both cases the simplified strain energy functional is written as

$$\begin{aligned}
\mathcal{A}(\mathbf{u},\mathbf{u}) &= \mathcal{A}_m(\mathbf{u},\mathbf{u}) + \mathcal{A}_s(\mathbf{u},\mathbf{u}) + \mathcal{A}_b(\mathbf{u},\mathbf{u}) \\
&= \int_\Gamma (\,\beta_{11}^2 + 2\beta_{12}^2 + \beta_{22}^2\,)\,dxdy + \frac{1}{2}\int_\Gamma (\rho_1^2 + \rho_2^2)\,dxdy \\
&+ \frac{d^2}{12}\int_\Gamma (\,\kappa_{11}^2 + 2\kappa_{12}^2 + \kappa_{22}^2\,)\,dxdy.
\end{aligned} \tag{6.3}$$

We name this simplified shell model as the *mathematical shell model*. This has thus two variants, the 5-field model where $\mathbf{u} = (u, v, w, \theta, \psi)$ where the strains are defined by Eq. (6.1) and the 3-field model where $\mathbf{u} = (u, v, w)$, the membrane strains are as in Eq. (6.1), the transverse shear strains vanish, and the bending strains are defined by Eq. (6.2).

In the mathematical shell models, the material parameters of the original shell problem are thus scaled off, and the geometry of the shell is reduced to four constant parameters a, b, c, d that define the curvature tensor of the midsurface and the thickness of the shell. The model may be viewed as a scaled local approximation of the original shell problem around a given point P of the midsurface. Following the usual geometric classification of surfaces, we call the shell *elliptic/parabolic/hyperbolic* when, respectively, $ab - c^2 > 0$ / $ab - c^2 = 0$ / $ab - c^2 < 0$.

7 Shell deformation states

When the load functional in the 3D energy expression (2.1) is given, the *shell deformation* is defined as the displacement field $\mathbf{u}$ that minimizes the energy over the *energy space* $\mathcal{U}$ under the assumed *kinematic constraints* (essential boundary conditions) on $\partial\Omega$. In the 3D formulation, the energy space is the Sobolev space $\mathcal{U} = H^1(\Omega)^3$, and the Euler equations of the energy minimization problem are the usual 3D equilibrium equations of linear elasticity. We now make use of the classical shell theory to get some basic unerstanding of the mathematical nature of shell deformations. To this end, we use the simplest shell model at hand, the mathemathical 3-field shell model as defined above.

In the 3-field models of classical shell theory, the energy functional (2.1) is defined in terms of the displacement field $\mathbf{u} = (u, v, w)$ that takes values in the Sobolev space (energy space) $\mathcal{U} = H^1(\Gamma) \times H^1(\Gamma) \times H^2(\Gamma)$. We will assume that the load functional is given as

$$\mathcal{L}(\mathbf{u}) = \int_\Gamma (f_1\, u + f_2\, v + f_3\, w)\, d\Gamma,$$

as corresponding to a given distributed surface traction $\mathbf{f} = (f_1, f_2, f_3)$ along Γ. (In the original 3D problem, the traction may act, e.g., on the outer surface of the shell at $z = d/2$.) In the mathematical 3-field model, the Euler equations of the energy minimization are then written as

$$\begin{cases} -\dfrac{\partial \beta_{11}}{\partial x} - \dfrac{\partial \beta_{12}}{\partial y} = f_1, \\ -\dfrac{\partial \beta_{12}}{\partial x} - \dfrac{\partial \beta_{22}}{\partial y} = f_2, \\ a\beta_{11} + b\beta_{22} + c\beta_{12} + \dfrac{d^2}{12}\Delta^2 w = f_3, \end{cases} \tag{7.1}$$

where β_{ij} are defined as in Eq. (6.1). We call these the *mathematical shell equations*.

In the mathematical shell equations (7.1) the principal parameter is the visible parameter d. This may be considered equal to the effective (dimesnionless) thickness $t_{\text{eff}} = d/L$

when L is chosen as the length unit of x, y. When the effective thickness is small, system (7.1) may be viewed as a singular perturbation of the corresponding *asymptotic* system obtained by setting $d = 0$. In the energy formulation this corresponds to setting $\rho_i = \kappa_{ij} = 0$ when minimizing the energy, i.e., the membrane energy is assumed to be the only form of deformation energy. This is a shell model in itself, referred to as the (shell) *membrane theory* in classical shell theory. In this model, one may aliminate w from the last equation in Eq. (7.1) so as to obtain a system of PDE's for u, v. This system is elliptic/parabolic/hyperbolic as corresponding to the geometry of the shell, so the shell geometry has a strong effect on the mathematical nature of the membrane theory. The precise formulation of the membrane theory, as well as the regularity theory of the solutions, are actually rather delicate mathematical problems, see [11, 12, 13] for a survey over a specific set of shell problems.

When the deformation of the shell varies smoothly in a length scale L and approaches the membrane-theory solution of the same nature when $d \to 0$ and the load is fixed, the deformation is called *membrane-dominated*. This is the first of the two main *deformation states* of a shell. For example, the deformation of a closed eggshell under a smooth loading is of this type. When the shell is not closed but has a boundary where some boundary conditions a imposed (say, piece of a pipe with clamped ends), the pure membrane-dominated behavior is less common, because a *boundary layer* typically appear at the boundary, and this does not obey the laws of membrane theory. Similar phenomena appear when the loading is not smooth. Even in such situations, the membrane-dominated deformation can still be considered as one *component* of the deformation. This component can be isolated from the layer by modifying the boundary conditions, so that the deformation is decomposed in two parts, the membrane-dominated part and the layer part. In this way the whole shell problem may be thought of being split in (two or more) subproblems in such a way that each 'feature' of the deformation is obtained as a solution to a single subproblem. Such a splitting, though only imagined, helps also to understand the finite element algorithms, see below.

The second main deormation state of the shell is a *bending-dominated* state. This can occur if the kinematic constraints at the boundary of the shell are sufficiently weak so as to allow a pure bending, i.e., a deformation $\mathbf{u}$ such that

$$\beta_{11}(\mathbf{u}) = \beta_{12}(\mathbf{u}) = \beta_{22}(\mathbf{u}) = 0. \tag{7.2}$$

If the load is able to exite such deformations, the solution at small values of d is obtained approximately my minimizing the energy this time under the constraints $\beta_{ij} = \rho_i = 0$. This is another asymptotic theory of shell deformations, known as the *inextensional theory*. A smooth, nearly inextensional deformation is called beding-dominated. Such a deformation occurs, e.g., on a piece of paper bent and clued to a cylindrical shell with free ends. Only for or special loads, such as uniform normal pressure, the deformation is membrane-dominated.

In the bening-dominated deformation state the deformation scales like $\mathbf{u} \sim d^{-2}$ as the load is fixed and d varies. The shell subject to such a deformation is thus rather 'soft' or 'sensitive' as compared with the membrane-dominated case. By eliminating u, v from

contstaints (7.2) one gets the equation

$$b\,\frac{\partial^2 w}{\partial x^2} + a\,\frac{\partial^2 w}{\partial y^2} - 2c\,\frac{\partial^2 w}{\partial x \partial y} = 0. \tag{7.3}$$

This equation is elliptic/parabolic/hyperbolic in the corresponding geometric categories of the shell, so the effect of the shell geometry is seen also in the inextensional theory. In view Eq. (5.5), Eq. (7.3) holds also in case of a cylindrical shell, with $a = c = 0$, $b = 1/R$. Thus the the pure bending of a cylindrical shell is such that w (u and v as well) is linear in the axial direction.

8 Boundary layers

Boundary layers in shell deformations are special, exponentially decaying solutions to the homogneous linear elastic equations in the shell geometry. To understand the main layer phenomenaa, it is again sufficient to consider the 3-field mathematical shell model. Thus we look for the boundary layers as special solutions to the mathematical shell equations (7.1) where $f_1 = f_2 = f_3 = 0$. Exponentially decaying layer solutions to these equations may basically arise as 'hot spots' or *point layers* or as *line layers*. We will here only summarize some of the results from [17] where line layers generated by a straight line were considered.

Suppose the layer generator is the line $x = 0$. (This could be a boundary line or a line where the load is irregular.) We look for solutions to the homogeneeous Eqs. (7.1) that decay in x in the halfspace $x > 0$. Assuming a Fourier transform with respect to the y-variable, we can obtain as special solutions *Fourier layer modes* of the form

$$\mathbf{u}(x, y) = \mathbf{u}(0)\, e^{iky} e^{-sx}, \tag{8.1}$$

where $\operatorname{Re} s > 0$. Given $k \in \mathbb{R}$, the possible values of $s = s(k, d)$ are found as (complex) solutions to a linear eigenvalue problem. We look in particular for boundary layer solutions such that $\operatorname{Re} s(k, d) \to \infty$ as $d \to 0$. To this end, the dominant terms of the characteristic are found to be [17]

$$\tfrac{1}{12}\, d^2\, s^8 + b^2\, s^4 + 4c^2\, s^2 + a^2 + [\ldots] = 0. \tag{8.2}$$

The solutions of the desired type can then be expanded in terms of fractional powers of d as

$$s(d, k) = A_0(k)\, d^{-1/m} + A_1(k)\, d^{1/m} + A_2(k)\, d^{3/m} + ..,$$

where either $m = 2$, $m = 3$, or $m = 4$, depending on the values of a, b, c. To find the leading terms in these expansions, it suffices to set $[\ldots] = 0$ in Eq. (8.2). The result is as

follows.

$$\underline{\text{Case 1}} \quad b \neq 0: \qquad s(d,k) = \left(-\frac{12\,b^2}{d^2}\right)^{1/4} + \ldots$$

$$\underline{\text{Case 2}} \quad b = 0,\ c \neq 0: \qquad s(d,k) = \left(\frac{48\,c^2k^2}{d^2}\right)^{1/6} + \ldots$$

$$\underline{\text{Case 3}} \quad b = c = 0,\ a \neq 0: \qquad s(d,k) = \left(-\frac{12\,a^2k^4}{d^2}\right)^{1/8} + \ldots$$

Here Case 1 is the main shell layer mode that is possible in all shell geometries. In Cases 2 and 3, the curvature along the layer line vanishes, which is possible in a hyperbolic (Case 2) and parbolic (Case 3) shell geometries.

For each Fourier layer mode of the above type with $\operatorname{Re} s > 0$, the characteristic length scale of decay is $L = 1/\operatorname{Re} s$. If we assume that $a, b, c \sim R^{-1}$ and $k \sim R^{-1}$, then the effective thickness $t_{\text{eff}} = d/L$ for the above three layer modes is of the order

$$t_{\text{eff}} \sim \begin{cases} (d/R)^{1/2}, & \text{Case 1,} \\ (d/R)^{2/3}, & \text{Case 2,} \\ (d/R)^{3/4}, & \text{Case 3.} \end{cases}$$

Since $t_{\text{eff}} \to 0$ when $d/R \to 0$, we conclude that the dimension performs well in the length scales of the layer modes when d/R is small. Thus we may expect that the layer modes obtained as solutions to the mathematical shell equations represent actual 3D phenomena.

When passing from the 3-field model to the 5-field version of the mathematical shell model, the above three layer modes remain essentially unchanged, while a new short-range layer mode in the length scale $L \sim d$ appears, with the corresponding root of the characteristic equation (now of degree 10) expanded as [17]

$$s(d,k) = -\frac{6}{d} + \ldots$$

Since the leading term does not depend oon the curvature parameters, the layer is present also in a flat plate/membrane problem where $a = b = c = 0$. The dimension reduction actually loses its validity in the range of this layer mode, since the effective for the characteristic length scale of the layer is $t_{\text{eff}} \sim 1$. Anyhow, the mere presence of the short-range layer in the 5-field model does indicate a true phenomenon in the original 3D formulation of the problem.

9 Shell deformation modes and the 3D FEM

To understand the performance of a finite element algorithm when modelling a shell deformation, it is necessary to split the problem in subproblems so that in each subproblem, the solution has a well defined, characteristic behavio from the numerical modelling point

of view. When isolating such subproblems, or deformation modes, classical shell theory is of great help. In the previous sections we have outlined the program of analysis that is needed as a basis of the mathematical understanding of finite element models of shell deformations. We should underline once more that we are speaking of understanding, not of programming. In the programming, the original 3D (or reduced 3D) formulation of the shell progrem is sufficient. Classical shell 2D theories are not needed.

The starting point of the finite element error analysis thus the expansion of the exact 3D solution of the shell problem as

$$\mathbf{u} = \mathbf{u}_1 + \mathbf{u}_2 + \dots \tag{9.1}$$

This corresponds to splitting the load functional as

$$\begin{aligned} \mathcal{L}(\mathbf{v}) = \mathcal{A}(\mathbf{u}, \mathbf{v}) &= \mathcal{A}(\mathbf{u}_1, \mathbf{v}) + \mathcal{A}(\mathbf{u}_2, \mathbf{v}) + \dots \\ &= \mathcal{L}_1(\mathbf{v}) + \mathcal{L}_2(\mathbf{v}) + \dots \end{aligned} \tag{9.2}$$

The functionals $\mathcal{L}_i$ may be viewed as *generalized loads* in the subproblems. Of course, to actually find such loads would reguire the exact solution of the whole set of subproblems and hence the original shell problem. The splitting is thus an imagined one only – it is not done in practice.

The variational formulation of the original shell problem is stated as: Find $\mathbf{u} \in \mathcal{U}$ satisfying the kinematic constraints of the problem and such that

$$\mathcal{A}(\mathbf{u}, \mathbf{v}) = \mathcal{L}(\mathbf{v}), \quad \mathbf{v} \in \mathcal{U}_0. \tag{9.3}$$

Here $\mathcal{U}$ is the 3D energy space and $\mathcal{U}_0$ is the subspace where the homogeneous versions of the kinematic constraints are imposed. When the splitting (9.1)–(9.2) is done, we must distribute also the kinematic constraints in the subproblems. Once this is done, the splitting (9.1)– (9.2) carries over to the variational formulation: Each $\mathbf{u}_i$ is defined as the solution to the subproblem where $\mathbf{u}_i \in \mathcal{U}$ satisfies the constraints of the subproblem and Eq. (9.3) with $\mathbf{u}_i$ replacing $\mathbf{u}$ and $\mathcal{L}_i$ replacing $\mathcal{L}$. One should note that the kinematic constraints in the subproblems are, in general, inhomogeneous even if the constraints of the original problem happen to be homogeneous.

The starting point of the finite element model is the variational formulation (9.3). We set up a finite element subspace $\mathcal{U}_h \in \mathcal{U}$ and we define the finite element solution as a function $\mathbf{u}_h \in \mathcal{U}_h$ satisfying the *interpolated* kinematic constraints at the boundary and such that

$$\mathcal{A}(\mathbf{u}_h, \mathbf{v}) = \mathcal{L}(\mathbf{v}), \quad \mathbf{v} \in \mathcal{U}_{h,0}. \tag{9.4}$$

The above splitting carries over to the finite element aproximation as well, so in the error analysis we may consider each subproblem separately. The motivation of the splitting actually comes here: If the splitting is done properly, we may be able to carry out the finite element error analysis in each subproblem using the special characteristics of the solution (deformation mode) $\mathbf{u}_i$. Indeed, the finite element error behavior turns out to be rather different for different deformation bounds as described in the previous sections, so

the splitting of the deformation into characteristic submodes is a must if we want sharp error analysis.

As a natural starting point of the finite element model we may consider an approximation that resembles the m-model described in Section 3. Here one discretizes the reference domain Ω' using a single layer of solid elements (prisms) and assumes a polynomial expansion of degree $m+1$ in z in each element. The error of such an approximation is naturally decomposed in two parts by writing

$$\mathbf{u} - \mathbf{u}_h = (\mathbf{u} - \mathbf{u}^{(m)}) + (\mathbf{u}^{(m)} - \mathbf{u}_h), \tag{9.5}$$

where $\mathbf{u}^{(m)}$ is the deformation of the shell according to the m-model. The first term in Eq. (9.5) is then the *modelling error*. To estimate this error term is a problem of shell theory. We concentrate here on the second error term, the bounding of which is a problem of numerical analysis. Here we assume moreover that $m = 1$.

When bounding the second term in Eq. (9.5) in case $m = 1$, we simplify the problem in the same manner as the 1-model itself was simplified above. Thus we interpret the 3D finite element scheme as 2D scheme within the 5-field mathematical shell model as described in Section 6. (As noted before, the 3-field model is inappropriate, since the constraints (5.8) are problematic to impose in numerical models.) The simplification of the finite element scheme from 3D to 2D is straightforward. What is leass clear is, whether all the essential characteristics of the scheme are preseved in the transition. Although this is by no means fully certified by our analysis, we can rely on the following conjecture that alone is sufficient to justify the simplification: Any essential numerical difficulties that are met in the finite element approximation within the mathematical shell model must also be met in the 3D finite element models of shells. In other words, the difficulties within the simplest shell model cannot be escaped by assuming more complex shell models as a starting point.

10 Finite element error bounds

Following the philosophy as set above we now look at the finite element approximations of shell deformations in more detail, assuming the mathematical 5-field model as the starting point. Then we have $\mathbf{u} = (u, v, w, \theta, \psi)$ and $\mathcal{U} = [H^1(\Gamma)]^5$ in Eq. (9.3), with the bilinear form $\mathcal{A}$ defined according to Eqs. (6.3) and (6.1). In the finite element formulation (9.4) we approximate each displacement component separately and in the same way, so that $\mathbf{u}_h = (u_h, v_h, w_h, \theta_h, \psi_h) \in [V_h]^5 = \mathcal{U}_h$, where $V_h \in H^1(\Gamma)$ is a standard 2D finite element space. We assume that V_h is associated to finite element subdivision of Γ into elements (say, triangles or quadrilaterals) of diameter at most h. The element shape functions are assumed to span all polynomials of given degree p, $p \geq 1$.

The above finite element scheme is an obvious interpretation of the discretized 1-model where the functions $u, v, w, \theta_1, \psi_1, \xi_1, \xi_2$ in Eq. (3.2) are approximated in the corresponding manner on the refernce domain ω. To carry out the error analysis of this scheme, we

choose the error indicator to be

$$e(\mathbf{u}) = \frac{|||\mathbf{u} - \mathbf{u}_h|||}{|||\mathbf{u}|||}\,, \tag{10.1}$$

where $|||\cdot|||$ is the *energy norm* defined by

$$|||\mathbf{u}||| = [\,\mathcal{A}(\mathbf{u},\mathbf{u})\,]^{1/2}. \tag{10.2}$$

The error indicator is thus the relative error in the energy norm. This may be viewed as the most favourable indicator, since the finite element method gives the best approximation in the energy norm (within the assumed kinematic constraints on the boundary).

When the exact solution is split according to Eq. (9.1), the finite element error according to indicator (10.1) is bounded accordingly as

$$e(\mathbf{u}) \le A_1\, e(\mathbf{u}_1) + A_2\, e(\mathbf{u}_2) + \ldots\,, \tag{10.3}$$

where $A_i = |||\mathbf{u}_i|||/|||\mathbf{u}|||$ is the relative amplitude of component $\mathbf{u}_i$ measured in the energy norm. We will assume the splitting to be defined so that the bound (10.3) is essentially sharp, i.e., so that the componentwise errors do not cancel. We have then reduced the problem of bounding $e(\mathbf{u})$ to that of bounding the componentwise errors $e(\mathbf{u}_i)$. Below we focus on such error terms, dropping henceforth the subindex i.

By definition, the finite element method is the best approximation method in the sense that

$$|||\mathbf{u} - \mathbf{u}_h||| \le |||\mathbf{u} - \tilde{\mathbf{u}}|||, \tag{10.4}$$

where $\tilde{\mathbf{u}}$ is the *interpolant* of $\mathbf{u}$ in $\mathcal{U}_h$. By the definition of the energy norm, the right side of this estimate can be bounded as

$$|||\mathbf{u} - \tilde{\mathbf{u}}||| \le C\|\mathbf{u} - \tilde{\mathbf{u}}\|_1\,, \tag{10.5}$$

where $\|\cdot\|_1$ stands for the norm of the Sobolev space $[H^1(\Gamma)]^5$. We define more generally the norm $\|\cdot\|_k$ of the Sobolev space $[H^k(\Gamma)]^5$ as

$$\|\mathbf{u}\|_k = \left\{\|u\|_k^2 + \|v\|_k^2 + \|w\|_k^2 + L^2\|\theta\|_k^2 + L^2\|\psi\|_k^2\right\}^{1/2}, \tag{10.6}$$

where

$$\|\phi\|_k^2 = \sum_{l=0}^{k}\sum_{i=0}^{l}\int_\Gamma \left(L^l\,\frac{\partial^l\phi}{\partial^i x\,\partial^{l-i}y}\right)^2 dxdy. \tag{10.7}$$

Here we have chosen L to be the length unit and taken into account that θ, ψ are dimensionless while the unit of u, v, w, x, y is L. Below we think of L as the characteristic length scale of the deformation component considered. In Eq. (10.5) and below, C stands for a generic constant of harmless size.

By standard finite element approximation theory, and by standard Sobolev-norm notation, we can bound the right side in Eq. (10.5) further as [1]

$$|||\mathbf{u} - \tilde{\mathbf{u}}||| \le C\left(\frac{h}{L}\right)^p |||\mathbf{u}|||_{p+1}\,. \tag{10.8}$$

We assume that $\mathbf{u}$ is sufficiently smooth so that the right side in Eq. (10.8) is finite, and define

$$Q_p(\mathbf{u}) = \frac{\|\mathbf{u}\|_{p+1}}{\|\mathbf{u}\|_1}\,. \tag{10.9}$$

This is a dimensionless constant that characterizes the regularity of $\mathbf{u}$ in the assumed length scale L.

Combining now Eqs. (10.4)–(10.9) we obtain the final error bound

$$e(\mathbf{u}) \le C\,Q_p(\mathbf{u})\,K(\mathbf{u})\,\left(\frac{h}{L}\right)^p, \tag{10.10}$$

where

$$K(\mathbf{u}) = \frac{\|\mathbf{u}\|_1}{|\!|\!|\mathbf{u}|\!|\!|}\,. \tag{10.11}$$

For typical shell deformation modes varying in a given length scale L as assumed, estimate (10.10) is essentially sharp, or at least cannot be improved without further assumptions on the finite element algorithm. The factor $K(\mathbf{u})$ in the error bound, as defined by Eq. (10.11), may then viewed as the unavoidable *error amplification factor* when approximating the field (component) $\mathbf{u}$ by means of standard finite elements of degree p. Since $|\!|\!|\mathbf{u}|\!|\!| \le C\|\mathbf{u}\|_1$, the factor $K(\mathbf{u})$ is at least of order unity. The remaining question then is, how big this factor actually is for typical shell deformations. This question is actually the basic reason for splitting the displacement field in subfields as assumed, since the question can only be given a definite answer for characteristic modes of deformation.

In the membrane-dominated deformation state the membrane energy $\mathcal{A}_m(\mathbf{u},\mathbf{u})$ dominates in Eq. (6.3). This doiminance may be characterized by the more quantitative estimate

$$\mathcal{A}_s(\mathbf{u},\mathbf{u}) + \mathcal{A}_b(\mathbf{u},\mathbf{u}) \le C\,(d/L)^2\,\mathcal{A}_m(\mathbf{u},\mathbf{u}). \tag{10.12}$$

This assumption together with the scaled *Korn inequality*

$$\|\mathbf{u}\|_1^2 \le C\,[\,\mathcal{A}_m(\mathbf{u},\mathbf{u}) + \mathcal{A}_s(\mathbf{u},\mathbf{u}) + (L/d)^2\,\mathcal{A}_b(\mathbf{u},\mathbf{u})\,] \tag{10.13}$$

implies that $|\!|\!|\mathbf{u}|\!|\!| \sim \|\mathbf{u}\|_1$. Hence one has $K(\mathbf{u}) \sim 1$ when approximating a membrane-dominated deformation that satisfies Eq. (10.12).

In the bending-dominated case one may assume more quantitatively that

$$\mathcal{A}_m(\mathbf{u}) + \mathcal{A}_s(\mathbf{u},\mathbf{u}) \le C\,\mathcal{A}_b(\mathbf{u},\mathbf{u}). \tag{10.14}$$

Together with Eq. (10.13) this implies that $|\!|\!|\mathbf{u}|\!|\!| \sim (d/L)\,\|\mathbf{u}\|_1$. Hence one has $K(\mathbf{u}) \sim L/d = 1/t_{\text{eff}}$ when approximating a bending-dominated deformation that satisifes Eq. (10.14)

When the field component $\mathbf{u}$ to be approximated is any of the boundary layer modes (Fourier modes) as described in Section 8, a more detailed asymptotic analysis is rquired to determine the size of the factor $K(\mathbf{u})$. This was carried out in [17], where it was shown

that also for the layer modes one has $K(\mathbf{u}) \sim L/d = 1/t_{\text{eff}}$, with L defind as the length scale of the decay of the layer mode.

Summarizing the above results we conclude that for standard finite elements of degree p, the error bound (10.10) holds with the regularity constant $Q_p(\mathbf{u})$ defined by Eqs. (10.9) and (10.6)–(10.7) and with the size of the error amplification factor $K(\mathbf{u})$ given for the different deformation types as

$$K(\mathbf{u}) \sim \begin{cases} 1 & \text{for a membrane-dominated deformation,} \\ 1/t_{\text{eff}} & \text{for a bending-dominated deformation,} \\ 1/t_{\text{eff}} & \text{for boundary layer modes.} \end{cases} \tag{10.15}$$

That error amplification by factor $K(\mathbf{u}) \sim L/d = 1/t_{\text{eff}}$ appears in many (in fact, most) deformation types of a shell is an unfortunate numerical phenomenon that is especially harmful in the lowest-order ($p = 1$) finite element approximation. In that case severe mesh overrefinement is needed to compensate for the effect when t_{eff} is small. At higher values of p, say with $p = 3$ or $p = 4$, the error amplification can be compensated by considerably milder mesh overrifenement, due to the fact that the factor K does not grow with p. The much better behavior of high-order elements has also been demonstrated by numerical experiments, see [2, 3, 6, 16]. The least conclusion from the error analysis and from the experiments is then that when modelling shell deformations by finite elements, software that offers the *free choice* of degree p is preferable.

11 The dream of the "shell element"

In the finite element modelling of shells, and more generally in the modelling of thin structures, there has been a long standing dream of dinding special low-order finite element formulations that avoid the parametric error growth, as detected above for various shell deformation types. The generic idea in such formulations is to modify the strain energy functional $\mathcal{A}$ numerically so that Eq. (9.4) gets rewritten as

$$\mathcal{A}_h(\mathbf{u}_h, \mathbf{v}) = \mathcal{L}(\mathbf{v}), \quad \mathbf{v} \in \mathcal{U}_{h,0}, \tag{11.1}$$

where the numerically modified functional $\mathcal{A}_h$ replaces $\mathcal{A}$. The modification causes another error term to be controlled, so the modification has to be done carefully. The aim of the modification is anyway to avoid the paramtric error growth when using the modified error indicator

$$e(\mathbf{u}) = \frac{|\!|\!| \mathbf{u} - \mathbf{u}_h |\!|\!|_h}{|\!|\!| \mathbf{u} |\!|\!|}, \tag{11.2}$$

where $|\!|\!| \cdot |\!|\!|_h$ is the modified energy (semi)norm defined by

$$|\!|\!| \mathbf{u} |\!|\!|_h = [\, \mathcal{A}_h(\mathbf{u}, \mathbf{u}) \,]^{1/2}. \tag{11.3}$$

Modified low-order finite formulation of the type (11.1) have been very successful on many thin-domain problems of linear elasticity. By now the theory based on the

error indicator (11.2)–(11.3) is also understood in most cases, cf. [14, 15] and the further references therein. In the shell modelling, however, there has been less success, regarding both practice and theory. Our recommendation concerning the numerical modelling of shells remains the same as in the previous section: One should go for standard finite elements of high order. Dreams may be left as dreams.

References

[1] D. Braess, *Finite elements*, Cambridge University Press, Cambridge, 1997.

[2] H. Hakula, Y. Leino and J. Pitkäranta, Scale resolution, locking, and high-order finite element modelling of shells, *Comput. Methods Appl. Mech. Engrg.* 133 (1996), 157-182.

[3] H. Hakula and J. Pitkäranta, Pinched shells of revolution: experiments on high-order FEM, in: Proceedings of the Third International Conference of Spectral and High-Order Methods (ICOSAHOM95), Houston, 1995, 193-201.

[4] W.T. Koiter, A consistent first approximation in the general theory of thin elastic shells, in: *Theory of Thin Shells*, Proceedings of the IUTAM Symposium, Delft 1959; North-Holland, Amsterdam, 1960.

[5] On the foundations of the linear theory of thin elastic shells, *Proc. Kon. Nederl. Akad. Wetensch.* B 73 (1970), 169-195.

[6] Y. Leino and J. Pitkäranta, On the membrane locking of $h - p$ finite elements in a cylindrical shell problem, *Int. J. Num. Methods Engrg.* 37 (1994), 1053-1070.

[7] A.E.H. Love, The small free vibrations and deformation of a thin elastic shell, *Philos. Trans. R. Soc.* A 179 (1888), 491-546.

[8] A.E.H. Love, *A Treatise on the Mathematical Theory of elasticity*, Fourth Edition, Dover 1944.

[9] P.M. Naghdi, Foundations of elastic shell theory, in *Progress in Solid Mechanics*, Vol. 4, I.N. Sneddon and R. Hill, eds., North-Holland 1963, 1-90.

[10] V.V. Novozhilov, *The Theory of Thin Shells*, Noordhoff, Leiden, 1959 (Transl.; first Russian ed: Oborongiz 1951).

[11] J. Piila and J. Pitkäranta, Characterization of the membrane theory of a clamped shell: the parabolic case, *Math. Models Methods Appl. Sci. (M^3AS)* 3 (1993), 417-442.

[12] J. Piila, Characterization of the membrane theory of a clamped shell: the elliptic case. *Math. Models Methods Appl. Sci.(M^3AS)* 4 (1994), 147-177.

[13] J. Piila, Characterization of the membrane theory of a clamped shell: the hyperbolic case. *Math. Models Methods Appl. Sci. (M^3AS)* 6 (1996), 169-194.

[14] J. Pitkäranta, The first locking-free plane-elastic finite element: historia mathematica, *Comput. Methods Appl. Mech. Engrg.* 190 (2000), 1323-1366.

[15] J. Pitkäranta, Mathematical and historical reflections on the lowest-order finite element models of thin structures, Preprint A449 (2002), Institute of Mathematics, Helsinki University of Technology, to appear in *Computers & Structures*.

[16] J. Pitkäranta, Y. Leino, O. Ovaskainen and J. Piila, Shell deformation states and the finite element method: A benchmark study of cylindrical shells, *Comput. Methods Appl. Mech. Engrg.* **133** (1996), 157-182.

[17] J. Pitkäranta, A.-M. Matache and C. Schwab, Fourier mode analysis of layers in shallow shell deformations, *Comput. Methods Appl. Mech. Engrg.* **190** (2001), 2943-2975.

[18] W.S. Wlassow (V.Z. Vlasov), *Allgemeine Schalenteorie und ihre Anwendung in der Technik*, Berlin Akademie 1958 (Transl.; original Russian vol.: Gostekhizdat. 1949).

COMPUTATIONAL MECHANICS OF METALLIC MATERIALS AT LARGE STRAINS

C. Teodosiu

LPMTM-CNRS, University Paris 13, 93430 Villetaneuse, France
teodosiu@lpmtm.univ-paris13.fr

Abstract. This lecture focuses on some advanced constitutive models, which describe the plastic behaviour of metallic materials under intense forming sequences that may involve complex strain paths and/or large accumulated strains. The presentation is basically limited to the cold deformation of polycrystalline materials used in sheet metal forming. Within this framework, it addresses both the mechanical modelling of the plastic anisotropy induced by the texture and microstructural evolutions and the finite-element implementation of such models.

1. INTRODUCTION

The plastic anisotropy of rolled and well-annealed metal sheets is mainly due to the crystallographic texture. On the other hand, their deformation-induced anisotropy is determined, to a large extent, by the interaction between preformed dislocation structures and the current loading. Both these types of plastic anisotropy have pronounced effects on the macroscopic behaviour of the sheet. For instance, Lian *et al.* [1] have shown that the necking behaviour in biaxial tension is very sensitive to the shape of the yield locus and, therefore, an accurate description of it is required. Zhang and Lee [2] and Yamamura *et al.* [3], among others, have pointed out the effect of the work-hardening on springback.

The plastic behaviour at large strains is commonly modelled by using phenomenological laws involving isotropic and/or non-linear hardening. However, experimental studies have revealed that sharp strain-path changes can induce a complex anisotropic hardening behaviour, especially in the case of steel sheets. On the other hand, TEM evidence (see, e.g. [4-6]) has convincingly shown that such effects are the macroscopic counterpart of the evolution of dislocation structures. Therefore, we shall also consider the hardening behaviour at large plastic strains, by using a dislocation-based microstructural model, initially proposed by Teodosiu and Hu [7-9], and further developed in [10, 11], which takes into account the plastic anisotropy induced by the evolution of dislocation structures, in addition to the isotropic and kinematic hardening. This model includes, besides a scalar measure of the isotropic hardening, three tensor-valued internal variables, which describe, respectively, the back-stress, the directional strength of dislocation structures and their polarity.

The incremental elastoplastic constitutive equations resulting from the rate models and their finite element implementation depend, of course, on the algorithms adopted for the time integration. For reasons of conciseness, we shall limit ourselves to mainly considering the tangent formulation of the incremental stress-strain relations, which is directly applicable to explicit time-marching schemes, referring to the available literature for more sophisticated, implicit algorithms.

2. CONSTITUTIVE MODELLING

We restrict ourselves to the cold deformation of metals and neglect any viscous effects on the work-hardening. The elastic strains are considered negligibly small as compared to the corresponding plastic strains. Only the contribution of the microstructural evolution and of the initial texture on the plastic behaviour is considered, as they are predominant at moderately large strains, while the influence of the texture evolution on the work-hardening is neglected.

In what follows, all tensor variables are denoted by bold-face symbols and their components, whenever used, are referred to a Cartesian orthogonal frame. The summation convention over repeated indices of such components is used throughout the paper. The superscripts T, S, and A denote the transpose, the symmetric part and the antisymmetric part of a second-order tensor, respectively. Let $\mathbf{A}, \mathbf{B}$ denote two second-order tensors and $\mathbf{S}$ a fourth-order tensor. We define the double-contracted tensor product between such tensors, denote by a colon, as

$$\mathbf{A} : \mathbf{B} = A_{ij} B_{ij}, \qquad (\mathbf{S} : \mathbf{A})_{ij} = S_{ijkl} A_{kl}, \qquad \mathbf{A} : \mathbf{S} : \mathbf{B} = A_{ij} S_{ijkl} B_{kl}.$$

We further define the norm of $\mathbf{A}$ as $\| \mathbf{A} \| = \sqrt{A_{ij} A_{ij}}$ and its direction, if $\mathbf{A}$ is non-zero, by $\mathbf{A} / \| \mathbf{A} \|$. Finally, the norm of $\mathbf{S}$ is defined by $\| \mathbf{S} \| = \sqrt{S_{ijkl} S_{ijkl}}$.

2.1 Kinematics of large elastoplastic deformations

Consider a body B, e.g. a metal sheet, at time t_0, and choose its configuration, say C_0, as reference configuration of B. Let C denote the current configuration of B at time t and let $\mathbf{x}_0$ and $\mathbf{x}$ denote the position vectors of a material point X in the configurations C_0 and C, respectively. When B undergoes a plastic deformation, it generally has not a global natural (i.e. stress-free) configuration. The local natural configuration $\hat{C}$ of a material neighbourhood $N(X)$ of the material point X is defined as the ideal configuration that $N(X)$ would assume if it were cut out and released from all constraints, the position of all crystal defects being kept constant, in order to preclude any plastic deformation. Let $\hat{C}_0$ be the local natural configuration of $N(X)$ obtained by the same procedure at time t_0. Clearly, when using an updated Lagrangian description (see Sect. 3), the time interval $[t_0, t]$ may be replaced by the current incremental lapse of time $[t, t + \Delta t]$.

Let $\mathrm{d}\mathbf{x}$, $\mathrm{d}\mathbf{x}_0$, $\mathrm{d}\hat{\mathbf{x}}$ and $\mathrm{d}\hat{\mathbf{x}}_0$ denote, respectively, the same infinitesimal material vector in the configurations $C, C_0, \hat{C}$ and $\hat{C}_0$, respectively (Fig. 1). We define the deformation gradient $\mathbf{F}$, the elastic distortions $\mathbf{F}^{\mathrm{e}}$ and $\mathbf{F}_0^{\mathrm{e}}$ at times t, respectively t_0, and the plastic distortion $\mathbf{F}^{\mathrm{p}}$ by the relations

$$\mathrm{d}\mathbf{x} = \mathbf{F}\, \mathrm{d}\mathbf{x}_0, \quad \mathrm{d}\mathbf{x} = \mathbf{F}^{\mathrm{e}}\, \mathrm{d}\hat{\mathbf{x}}, \quad \mathrm{d}\hat{\mathbf{x}} = \mathbf{F}^{\mathrm{p}}\, \mathrm{d}\hat{\mathbf{x}}_0, \quad \mathrm{d}\mathbf{x}_0 = \mathbf{F}_0^{\mathrm{e}}\, \mathrm{d}\hat{\mathbf{x}}_0. \qquad (2.1)$$

This leads to the following multiplicative decomposition of the deformation gradient $\mathbf{F}$ into elastic and plastic parts

$$\mathbf{F} = \mathbf{F}^{e}\ \mathbf{F}^{p}\ (\mathbf{F}_0^{e})^{-1}. \tag{2.2}$$

By time-differentiating this relation, it follows that the gradient $\mathbf{L}$ of the velocity field can be written as

$$\mathbf{L} = \dot{\mathbf{F}}\,\mathbf{F}^{-1} = \dot{\mathbf{F}}^{e}\,(\mathbf{F}^{e})^{-1} + \mathbf{F}^{e}\,\dot{\mathbf{F}}^{p}\,(\mathbf{F}^{p})^{-1}(\mathbf{F}^{e})^{-1}. \tag{2.3}$$

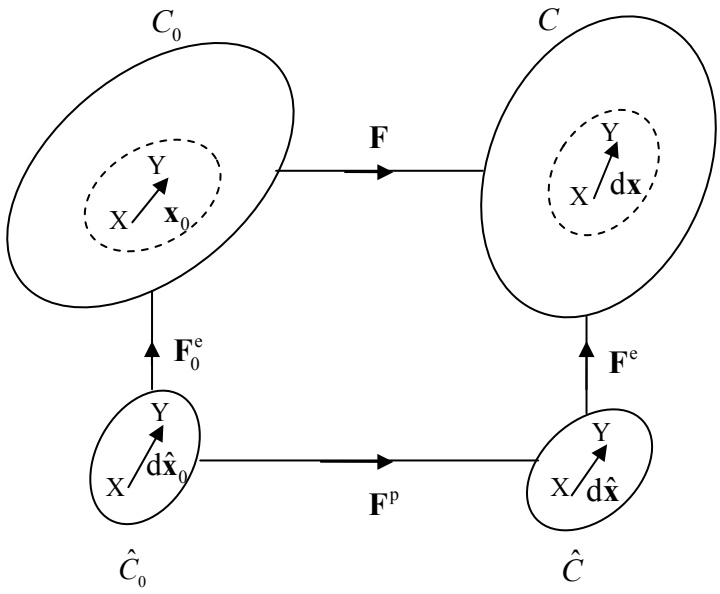


Figure 1. On the multiplicative decomposition of the elastoplastic deformation.

For cold deformation of metals, the elastic part of $\mathbf{F}^{e}$ involves strains that are a minute fraction of unity, but possibly large rotations. That is why, by using the polar decomposition of $\mathbf{F}^{e}$, we shall write $\mathbf{F}^{e} = (\mathbf{1} + \mathbf{e})\,\mathbf{R}$, where $\mathbf{1}$ is the unit second-order tensor, $\mathbf{e}$ is a symmetric tensor of small elastic strains ($\mathbf{e} = \mathbf{e}^{T}$, $\|\mathbf{e}\| \ll 1$), and $\mathbf{R}$ is the rotation tensor ($\mathbf{R}^{T}\mathbf{R} = 1$). By introducing this expression of $\mathbf{F}^{e}$ into (2.2) and neglecting terms of second order in $\|\mathbf{e}\|$, we obtain

$$\mathbf{L} = \dot{\mathbf{R}}\,\mathbf{R}^{T} + \overset{o}{\mathbf{e}} + \mathbf{R}\,\dot{\mathbf{F}}^{p}\,(\mathbf{F}^{p})^{-1}\mathbf{R}^{T}, \tag{2.4}$$

where

$$\overset{o}{\mathbf{e}} = \dot{\mathbf{e}} - \dot{\mathbf{R}}\,\mathbf{R}^{T}\mathbf{e} + \mathbf{e}\,\dot{\mathbf{R}}\,\mathbf{R}^{T}$$

denotes the objective time-derivative of $\mathbf{e}$, calculated with the spin $\dot{\mathbf{R}}\mathbf{R}^{\mathrm{T}}$. The symmetric and antisymmetric parts of (2.3) give the strain rate tensor $\mathbf{D}$ and the spin $\mathbf{W}$, respectively, as

$$\mathbf{D} = \overset{\circ}{\mathbf{e}} + \mathbf{D}^{\mathrm{p}} \qquad \mathbf{W} = \dot{\mathbf{R}}\mathbf{R}^{\mathrm{T}} + \mathbf{W}^{\mathrm{p}}, \tag{2.5}$$

where

$$\mathbf{D}^{\mathrm{p}} = \mathbf{R}\,\hat{\mathbf{D}}^{\mathrm{p}}\,\mathbf{R}^{\mathrm{T}}, \qquad \mathbf{W}^{\mathrm{p}} = \mathbf{R}\,\hat{\mathbf{W}}^{\mathrm{p}}\,\mathbf{R}^{\mathrm{T}} \tag{2.6}$$

are the plastic strain rate and the plastic spin, and $\hat{\mathbf{D}}^{\mathrm{p}}$ and $\hat{\mathbf{W}}^{\mathrm{p}}$ are the symmetric and antisymmetric parts of $\dot{\mathbf{F}}^{\mathrm{p}}\,(\mathbf{F}^{\mathrm{p}})^{-1}$, respectively.

The as yet undefined orientations of the local natural configurations $N(X)$ can be chosen so as to simplify the description of the material response to subsequent deformations. For single crystals, this can be achieved by making the average lattice orientation of $N(X)$ be the same throughout the motion [12-14]. Indeed, this renders homogeneous the description of the elastic response when neglecting the influence of the plastic deformation on the elastic moduli. In addition, since the plastic deformation does not change the average lattice orientation, this choice justifies calling plastic deformation the mapping of C_0 onto $\hat{C}_0$.

Mandel [15] has proposed a natural extension of these 'isoclinic natural configurations' to polycrystalline materials, by associating to each material neighbourhood $N(X)$ a preferred frame that rotates with a spin equal to the volume average of the spins of all grains in $N(X)$. Clearly, this formalism requires describing the initial texture of the metal by a sufficiently large number of grain orientations, following the evolution of these orientations throughout the deformation process, and deriving the macroscopic behaviour by a suitable micro-macro transition scheme. Although several procedures of this type are already available in the literature (see, e.g., [16-18]), their use remains limited to rather simple deformation processes. Therefore, we will limit ourselves in the following to a simplified approach, which can be considered as acceptable, for instance when simulating the sheet metal forming. Namely, since the work-hardening behaviour of polycrystalline materials at moderately large strains is not very sensitive to the choice of the spin $\dot{\mathbf{R}}\mathbf{R}^{\mathrm{T}}$, we shall simply assume that $\mathbf{W}^{\mathrm{p}} = \mathbf{0}$ in $(2.5)_2$, the rotation field $\mathbf{R}(t)$ being determined by the evolution equation

$$\dot{\mathbf{R}} = \mathbf{W}\mathbf{R}, \tag{2.7}$$

with the initial condition $\mathbf{R}(0) = \mathbf{R}_0$, where $\mathbf{R}_0$ is the orientation of the preferred frame at the beginning of the deformation process, e. g. the frame defined by the axes of plastic orthotropy of a rolled sheet. Moreover, since all crystal defects are corotational with the crystal lattice, it will be assumed that all tensor hardening variables turn with the same spin $\mathbf{W} = \dot{\mathbf{R}}\mathbf{R}^{\mathrm{T}}$, while their objective rates, denoted by a small superposed circle, will be of Jaumann type.

2.2 Hypoelastic behaviour

Assuming again that the elastic strains are small and that the axes of elastic and plastic anisotropy coincide, it may be shown that the time-differentiation of the hyperelastic constitutive equation leads to the hypoelastic form

$$\overset{\circ}{\boldsymbol{\sigma}} = \mathbf{c} : \overset{\circ}{\mathbf{e}} = \mathbf{c} : (\mathbf{D} - \mathbf{D}^{\mathrm{p}}), \tag{2.8}$$

where $\mathbf{c}$ is the tensor of elastic constants, whereas the objective time derivatives $\overset{\circ}{\boldsymbol{\sigma}}$ and $\overset{\circ}{\mathbf{e}}$ are given by

$$\overset{\circ}{\boldsymbol{\sigma}} = \dot{\boldsymbol{\sigma}} - \mathbf{W}\boldsymbol{\sigma} + \boldsymbol{\sigma}\mathbf{W}, \qquad \overset{\circ}{\boldsymbol{\varepsilon}} = \dot{\boldsymbol{\varepsilon}} - \mathbf{W}\boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}\mathbf{W}. \tag{2.9}$$

2.3 Yield condition and flow rule

We assume that the yield condition has the form

$$\Phi = \bar{\sigma} - Y = 0, \tag{2.10}$$

where Y is the yield stress. The equivalent effective stress $\bar{\sigma}$ is given by the general quadratic function

$$\bar{\sigma}^2 = \mathbf{s} : \mathbf{M} : \mathbf{s}, \tag{2.11}$$

where $\mathbf{s} = \boldsymbol{\sigma}' - \mathbf{X}$ is the effective deviatoric stress tensor, $\boldsymbol{\sigma}'$ is the deviator of the Cauchy stress tensor, $\mathbf{X}$ denotes the back-stress, and $\mathbf{M}$ is a fourth-order tensor characterizing the texture anisotropy. Since $\mathbf{s}$ is deviatoric and symmetric, it may be assumed without restriction of generality that $\mathbf{M}$ is a fully symmetric tensor of fourth-order that is traceless in both the first and the second pair of indices, i. e.

$$M_{ijkm} = M_{jikm} = M_{kmij}, \qquad M_{iikm} = 0. \tag{2.12}$$

In fact, $\mathbf{M}$ should be added to the set of internal variable, too. However, since we neglect the texture evolution, we simply assume that the laminated sheet is initially orthotropic and remains so during the deformation. Then, the orthotropy axes, which are supposed to coincide initially with the fixed Cartesian axes x_i, are subjected to the time-dependent rotation $\mathbf{R}$, while the evolution of $\mathbf{M}$ is given by

$$M_{ijkm} = R_{in}\, R_{jp}\, R_{kq}\, R_{ms}\, \hat{M}_{npqs}. \tag{2.13}$$

Alternatively, since the components of $\mathbf{M}$ in the corotational frame remain constant and equal to those of $\hat{\mathbf{M}}$ in the fixed Cartesian frame, we may calculate the equivalent effective stress by

$$\bar{\sigma}^2 = \hat{M}_{ijkm}\, \hat{s}_{ij}\, \hat{s}_{km}, \tag{2.14}$$

where

$$\hat{s}_{ij} = R_{in}\, R_{jp}\, s_{np} \tag{2.15}$$

are the components of the effective stress in the (current) orthotropy frame. In particular, if the initial plastic anisotropy can be described by Hill's quadratic yield condition, Eq. (2.11) may be written in the form

$$\bar{\sigma}^2 = F(\hat{s}_{22} - \hat{s}_{33})^2 + G(\hat{s}_{33} - \hat{s}_{11})^2 + H(\hat{s}_{11} - \hat{s}_{22})^2 + 2L\,\hat{s}_{23}^2 + 2M\,\hat{s}_{13}^2 + 2N\,\hat{s}_{12}^2, \tag{2.16}$$

where *F, G, H, L, M* and *N* are material constants.

The plastic strain rate is given by the associated flow rule

$$\mathbf{D}^{\mathrm{p}} = \dot{\lambda}\left(\frac{\partial \Phi}{\partial \boldsymbol{\sigma}}\right)^{\mathrm{S}} = \frac{\dot{\lambda}}{\bar{\sigma}}\,\mathbf{M} : \mathbf{s}, \tag{2.17}$$

where $\dot{\lambda}$ is the plastic multiplier and a superposed dot indicates the time differentiation. The equivalent plastic strain rate $\dot{\bar{\varepsilon}}^{\mathrm{p}}$ is defined as the power conjugate of $\bar{\sigma}$, i.e.

$$\bar{\sigma}\,\dot{\bar{\varepsilon}}^{\mathrm{p}} = \mathbf{s} : \mathbf{D}^{\mathrm{p}}. \tag{2.18}$$

Introducing (2.17) into this relation and considering (2.11), it follows that $\dot{\bar{\varepsilon}}^{\mathrm{p}} = \dot{\lambda}$. The equivalent plastic strain is defined by

$$\bar{\varepsilon}^{\mathrm{p}} = \int_0^{\mathrm{t}} \dot{\bar{\varepsilon}}^{\mathrm{p}}\, \mathrm{dt} = \int_0^{\mathrm{t}} \frac{\mathbf{s} : \mathbf{D}^{\mathrm{p}}}{\bar{\sigma}}\, \mathrm{dt}. \tag{2.19}$$

When the initial texture is isotropic, we may replace the tensor $\mathbf{M}$ by

$$\mathbf{M} = (3/2)\,\mathbf{I}', \tag{2.20}$$

where $\mathbf{I}'$ is the unit fourth-order tensor in the applications of the space of symmetric and deviatoric tensors onto itself, and has the components

$$I'_{ijkm} = \frac{1}{2}\left(\delta_{ik}\,\delta_{jm} + \delta_{im}\,\delta_{jk}\right) - \frac{1}{3}\,\delta_{ij}\,\delta_{km}.$$

Introducing (2.20) into (2.11) and taking into account that $\mathbf{I}' : \mathbf{s} = \mathbf{s}$, yields the expression of the generalized von Mises equivalent stress

$$\bar{\sigma}^2 = \frac{3}{2}\,\mathbf{s} : \mathbf{s}. \tag{2.21}$$

It may be shown that in a uniaxial tensile tes, this equivalent stress reduces to the effective tensile stress, which explains the choice of the normalizing factor 3/2 in (2.20). Furthermore, replacing $\mathbf{M}$ by (2.20) into the flow rule (2.17) gives the relation

$$\mathbf{D}^{\mathrm{p}} = \frac{3\dot{\lambda}}{2\bar{\sigma}}\,\mathbf{s}, \tag{2.22}$$

which shows that the strain rate tensor and the effective stress tensor are now coaxial. It is worth noting that, although in this case the yield criterion is isotropic with respect to $\mathbf{s}$, it is still anisotropic with respect to $\boldsymbol{\sigma}'$, whenever the back-stress is non-zero.

2.4 Isotropic hardening combined with saturated kinematic hardening

One of the most common law used for the isotropic hardening is the Swift law defined by

$$Y = C\,(\varepsilon_0 + \bar{\varepsilon}^{\mathrm{p}})^n, \tag{2.23}$$

where C, ε_0 and n are material parameters. The initial value of the yield stress is given by the relation $Y_0 = C\,\varepsilon_0^n$. The Swift law is adequate for describing the behaviour of materials that exhibit a non-saturated isotropic hardening up to rupture.

Another frequent description of the isotropic hardening is given by the Voce law:

$$Y = Y_0 + R, \qquad R = C_{\mathrm{R}}\,(R_{\mathrm{sat}} - R)\dot{\lambda}, \qquad R(0) = 0, \tag{2.24}$$

where C_{R} and R_{sat} are material parameters, Y_0 is the initial yield limit, while the evolution of R describes the isotropic hardening. It is worth noting that by integrating the evolution equation $(2.24)_2$ with the initial condition $(2.24)_3$, the Voce law can be also written in the alternative algebraic form

$$Y = Y_0 + R_{\mathrm{sat}}\left[1 - \exp(-C_{\mathrm{R}}\,\bar{\varepsilon}^{\mathrm{p}})\right],$$

which shows that Y approaches asymptotically the value $Y_0 + R_{\mathrm{sat}}$ under monotonic loading. Therefore, the Voce law is adequate for describing the behaviour of materials that exhibit a saturated isotropic hardening before rupture.

The isotropic hardening can be combined with a kinematic hardening characterized by the evolution of the back-stress $\mathbf{X}$. We shall assume here that this evolution is governed by the saturation law that has been thoroughly investigated by Lemaître and Chaboche [19]:

$$\overset{\circ}{\mathbf{X}} = K\,\mathbf{D}^{\mathrm{p}} - \gamma\,\mathbf{X}\dot{\lambda}, \qquad \mathbf{X}(0) = \mathbf{X}_0, \tag{2.25}$$

where K and γ are material parameters and $\mathbf{X}_0$ is the initial value of $\mathbf{X}$. The ratio K/γ characterizes the saturation value of the kinematic hardening, while γ characterizes its rate of approaching the saturation.

A slightly different evolution equation for the back-stress proposed in the literature is

$$\overset{\circ}{\mathbf{X}} = C_{\mathrm{X}}\left[(X_{\mathrm{sat}}/\bar{\sigma})\,\mathbf{s} - \mathbf{X}\right]\dot{\lambda}, \qquad \mathbf{X}(0) = \mathbf{X}_0, \tag{2.26}$$

where X_{sat} characterizes the saturation value of $\mathbf{X}$, while C_{X} characterizes its rate of approaching the saturation. In the isotropic case, by considering (2.22), it is easily shown that the evolution equations (2.25) and (2.26) coincide, provided that $\gamma = C_{\mathrm{X}}$ and $K = (2/3)\,C_{\mathrm{X}}\,X_{\mathrm{sat}}$. On the other hand, when no rotation takes place, Eqs. (2.25) predict that

the back-stress tends to be coaxial with a constant plastic strain rate $\mathbf{D}^p$, while Eqs. (2.26) predict that the back-stress tends to be coaxial and opposite with a constant deviatoric stress $\boldsymbol{\sigma}'$, and these two predictions are different when the plastic anisotropy is significant. The latter conjecture seems more plausible from the physical point of view, because it corresponds to the very definition of the back stress as opposing the applied shear stress on each slip system. Therefore, we shall generally prefer using the evolution equations (2.26), in particular for the dislocation-based model that will be presented in the following subsection.

The mixed models obtained by combining one of the equations (2.23) and (2.24) with one of the equations (2.25) and (2.26) is adequate for materials exhibiting both isotropic hardening and a rather pronounced Bauschinger behaviour. Such a model involves 5 material parameters, which can be identified e.g. by using a uniaxial tensile test along the rolling direction, and monotonic and Bauschinger simple shear tests along the rolling direction.

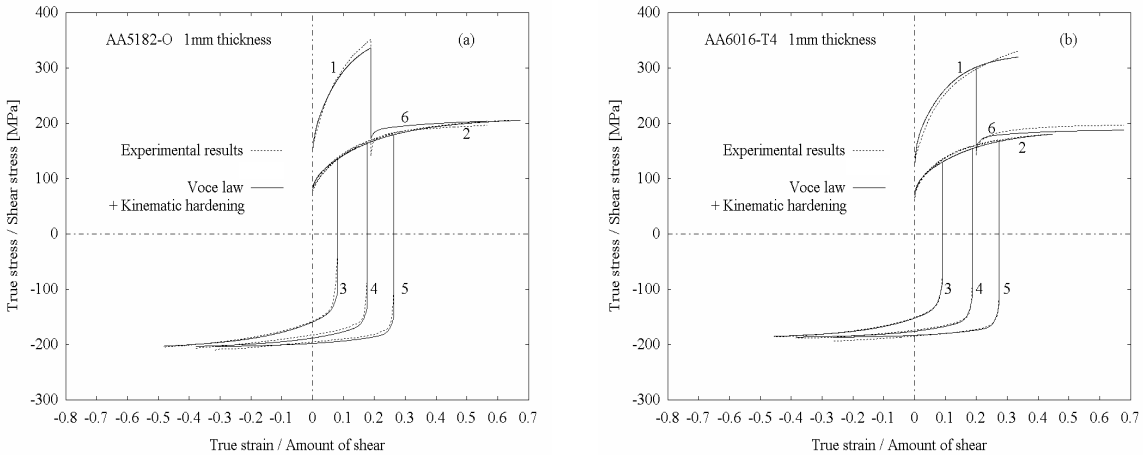


Figure 1. Comparison of mechanical tests with the prediction of the model combining isotropic hardening (Voce law) with kinematic hardening for aluminium alloys (a) AA5182-O and (b) AA6016-T4. (1) Uniaxial tensile test. (2) Monotonic simple shear test. (3), (4), (5) Bauschinger simple shear tests after 10%, 20% and 30% amount of shear in the forward direction. (6) Orthogonal test: simple shear in the rolling direction, following a 20% true tensile strain in the same direction (after [11]).

As an illustration, let us consider the description of the mechanical behaviour of the sheets of aluminium alloys AA5182-O and AA6016-T4. Several uniaxial tensile tests and simple shear tests at different orientations show that these materials exhibit a weak planar anisotropy in the flow stress, and a pronounced saturation of the flow stress during monotonic loadings. The Bauschinger effect is rather small and no work-hardening stagnation occurs during the reversed deformation of Bauschinger tests. After an orthogonal strain-path change, no cross-hardening or softening effects are detected in the alloy AA5182-O, while the alloy AA6016-T4 presents an increase in the yield stress, which is not followed, however, by any softening effect.

Figure 1 shows that this behaviour can be quite satisfactorily described by a combined model of isotropic hardening (Voce law) and saturating kinematic hardening. Indeed,

although the Bauschinger effect is relatively small, we observe that the match of experimental data is significantly improved when introducing the back-stress. The values of the hardening parameters for the alloy AA5182-O are: Y_0 = 148.5 MPa, C_R = 9.7, R_{sat} = 192.4 MPa, K = 2647 MPa, γ = 152.7 and those of the Hill parameters are F = 0.652, G = 0.570, H = 0.430, N = 1.61. The values of the hardening parameters for the alloy AA6016-T4 are: Y_0 = 124.2 MPa, C_R = 9.5, R_{sat} = 167.0 MPa, K = 3409 MPa, γ = 146.5 and those of the Hill parameters are F = 0.587, G = 0.590, H = 0.410, N = 1.27.

2.5 A dislocation-based microstructural model

In this section we will present the microstructural model proposed by Teodosiu and Hu [7-9], with its subsequent refinements introduced in [10, 11], which is intended to describe the plastic behaviour of some metallic rolled sheets at large strains and under complex strain-path changes. We shall limit ourselves to analyse the results obtained for the IF steel DC06 and the dual phase steel DP600, referring for other results concerning the high-strength steel HSLA340 and the aluminium alloys AA5182-O and AA6016-T4 to the report [11].

The hardening of the material is described by four internal state variables denoted by R, $\mathbf{X}$, $\mathbf{P}$ and $\mathbf{S}$. The tensor variables $\mathbf{S}$ and $\mathbf{P}$ are associated, respectively, with the directional strength of planar dislocation sheets and with their polarity. $\mathbf{S}$ is a fourth-order tensor and has the dimension of stress, while $\mathbf{P}$ is a second-order tensor and has no dimension. $\mathbf{X}$ is a second-order deviatoric tensor and represents a kind of generalized back-stress, which is intended to describe the rapid changes in the flow stress following a sharp change in the direction of the strain rate. Finally, R is a scalar variable and describes the isotropic work-hardening associated with the randomly distributed dislocations. For well-annealed materials, the initial values of all these internal variables are set equal to zero. On the contrary, for a predeformed material as a cold-rolled sheet, these initial values should be incorporated in the identification procedure of the model.

The yield condition is still written in the form (2.10), but the yield stress is defined now by the relation

$$Y = Y_0 + R + f\,\|\mathbf{S}\|, \tag{2.27}$$

where Y_0 is the initial yield stress, f is a material parameter, whereas the terms R and $f\,\|\mathbf{S}\|$ describe, respectively, the contributions of the randomly distributed dislocations and of the dislocation structures to the isotropic hardening.

The definitions of the equivalent effective stress, the associated flow rule and the evolution equations of the hardening variables R and $\mathbf{X}$ conserve the forms (2.11), (2.17), $(2.23)_2$ and $(2.26)_1$, respectively. However, the parameter X_{sat} in Eq. $(2.26)_1$ is no longer assumed constant, but considered as a function of the dislocation structures, via the internal state variable $\mathbf{S}$. More precisely, it is assumed that

$$X_{sat} = X_0 + (1-f)\sqrt{S_D^2 + r\,\mathbf{S}_L^2}, \tag{2.28}$$

where

$$S_{\mathrm{D}} = \mathbf{A} : \mathbf{S} : \mathbf{A} \tag{2.29}$$

is the strength of the dislocation structures associated with the currently active slip systems,

$$\mathbf{A} = \frac{\mathbf{D}^{\mathrm{p}}}{\left\| \mathbf{D}^{\mathrm{p}} \right\|} \tag{2.30}$$

is the current direction of the strain rate tensor, and $\mathbf{S}_{\mathrm{L}}$ is the part of $\mathbf{S}$ associated with the latent slip systems, which is defined by

$$\mathbf{S}_{\mathrm{L}} = \mathbf{S} - S_{\mathrm{D}}\, \mathbf{A} \otimes \mathbf{A}. \tag{2.31}$$

TEM experimental evidence [4-6] strongly suggests that dislocation structures associated with the current direction $\mathbf{A}$ of the strain rate tensor evolve quite differently from the latent dislocation structures, which explains the decomposition of $\mathbf{S}$ into S_{D} and $\mathbf{S}_{\mathrm{L}}$. Moreover, for some metallic materials, e.g. the ferritic steels, two physical mechanisms may intervene immediately after an orthogonal strain-path change: either the partial disintegration of the latent dislocation structures or their softening after being sheared microbands associated with the newly activated slip systems. Both these mechanisms reduce the intensity of the latent part $\mathbf{S}_{\mathrm{L}}$ of the dislocation structures, whereas the part S_{D} associated with the currently active slip systems will increase. Hence, the following evolution equations of S_{D} and $\mathbf{S}_{\mathrm{L}}$ are adopted:

$$\overset{\circ}{\mathbf{S}}_{\mathrm{L}} = -C_{\mathrm{SL}} \left(\frac{\left\| \mathbf{S}_{\mathrm{L}} \right\|}{S_{\mathrm{sat}}} \right)^{\mathrm{n}} \mathbf{S}_{\mathrm{L}}\, \dot{\lambda}, \tag{2.32}$$

$$\dot{S}_{\mathrm{D}} = C_{\mathrm{SD}} \left[\left(S_{\mathrm{sat}} - S_{\mathrm{D}} \right) g - S_{\mathrm{D}}\, h \right] \dot{\lambda}, \tag{2.33}$$

where C_{SD} and C_{SL} characterize the saturation rates of S_{D} and $\mathbf{S}_{\mathrm{L}}$, respectively, S_{sat} denotes the saturation value of S_{D}, g is a scalar function describing the influence of the polarity of the planar dislocation structures, and h is a scalar function describing the slight variation of S_{D} at the beginning of the reversed deformation in a Bauschinger test and vanishing thereafter. Specifically, by denoting

$$P_{\mathrm{A}} = \mathbf{P} : \mathbf{A}, \qquad X_{\mathrm{A}} = \mathbf{X} : \mathbf{A}, \tag{2.34}$$

the functions g and h may be expressed as

$$g = \begin{cases} 1 - \dfrac{C_{\mathrm{P}}}{C_{\mathrm{SD}} + C_{\mathrm{P}}} \left| \dfrac{S_{\mathrm{D}}}{S_{\mathrm{sat}}} - P_{\mathrm{D}} \right| & \text{if } P_{\mathrm{D}} \geq 0 \\ (1 + P_{\mathrm{D}})^{n_{\mathrm{P}}} \left(1 - \dfrac{C_{\mathrm{P}}}{C_{\mathrm{SD}} + C_{\mathrm{P}}} \dfrac{S_{\mathrm{D}}}{S_{\mathrm{sat}}} \right) & \text{otherwise} \end{cases} \tag{2.35}$$

$$h = \frac{1}{2} \left(1 - \frac{X_{\mathrm{A}}}{\mathrm{X}_{\mathrm{sat}}} \right) \tag{2.36}$$

The initial values of S_{D} and $\mathbf{S}_{\mathrm{L}}$ may be obtained from the initial values $\mathbf{S}_0$ and $\mathbf{A}_0$ of $\mathbf{S}$, respectively $\mathbf{A}$, by using Eqs. (2.29) and (2.31), i.e.

$$S_{\mathrm{D}}(0) = \mathbf{A}_0 : \mathbf{S}_0 : \mathbf{A}_0, \qquad \mathbf{S}_{\mathrm{L}}(0) = \mathbf{S}_0 - S_{\mathrm{D}}(0)\, \mathbf{A}_0 \otimes \mathbf{A}_0. \tag{2.37}$$

It should be mentioned, however, that only $\mathbf{S}$ is an internal variable, whereas its decomposition into S_{D} and $\mathbf{S}_{\mathrm{L}}$ is only a means of getting more physical insight into the evolution equations postulated for various parts of $\mathbf{S}$.

Finally the evolution law for $\mathbf{P}$ is assumed in the form

$$\overset{\circ}{\mathbf{P}} = C_{\mathrm{P}} (\mathbf{A} - \mathbf{P}) \dot{\lambda}, \tag{2.38}$$

which shows that, whatever the initial value of $\mathbf{P}$, it will tend to $\mathbf{A}$ if the direction of the strain rate remains unchanged for an amount of deformation that is sufficiently large with respect to $1/C_{\mathrm{P}}$.

In its extensive form, which is mainly used for mild and IF steels, the dislocation-based model involves 13 material parameters, namely Y_0, f, C_{R}, R_{sat}, C_{X}, X_0, C_{P}, C_{SD}, S_{sat}, n_{P}, C_{SL}, r, n. The somewhat simpler behaviour of other steels, which do not display a plateau in Bauschinger tests or a temporary work-softening after orthogonal strain-path changes, can be derived from the general model, by simply setting to zero some of the material parameters. This is the case, for example, of the dual phase DP600 presented in Fig. 2b.

Figure 2 depicts the behaviour of two steels. The IF steel, whose behaviour is illustrated by Fig. 2a, exhibits a work-hardening stagnation followed by a resumption of work-hardening during the reversed deformation in Bauschinger tests. The length of the plateau increases with the amount of forward shear, during the first strain path. A cross-hardening effect is also detected after an orthogonal strain-path change, consisting of a temporary work-hardening followed by work-softening. As shown in Fig. 2a, the dislocation-based microstructural model is able to completely predict these various features. The values of the hardening parameters for the steel DC06 are: Y_0 = 121.1 MPa, C_{R} = 31.9, R_{sat} = 90 MPa, C_{X} = 446, X_0 = 15.9 MPa, C_{SD} = 4, C_{SL} = 1.86, S_{sat} = 231.1 MPa, n = 0, n_{P} = 27.9, r = 1.5, f = 0.445, C_{P} = 5.5, whereas those of the Hill's parameters are F = 0.243, G = 0.297, H = 0.703, N = 1.20.

The behaviour of the dual phase steel DP600, depicted in Fig. 2b, is slightly different from that of IF steel. In particular, no cross-hardening effects are observed after an orthogonal strain path change, but the presence of a plateau is clearly noticed in Bauschinger tests. As shown in Fig. 2a, the dislocation-based model is still able to completely depict these different observed features, by setting some parameters equal to 0. The values of the hardening parameters for the steel DP600 are: Y_0 = 285 MPa, C_R = 37.6, R_{sat} = 110.8 MPa, C_X = 55.7, X_0 = 169.4 MPa, C_{SD} = 5.6, C_{SL} = 0, S_{sat} = 330.7 MPa, n = 0, n_P = 664.5, r = 0, f = 0.631, C_P = 0.54, whereas those of the Hill's parameters are F = 0.503, G = 0.559, H = 0.441, N = 1.49.

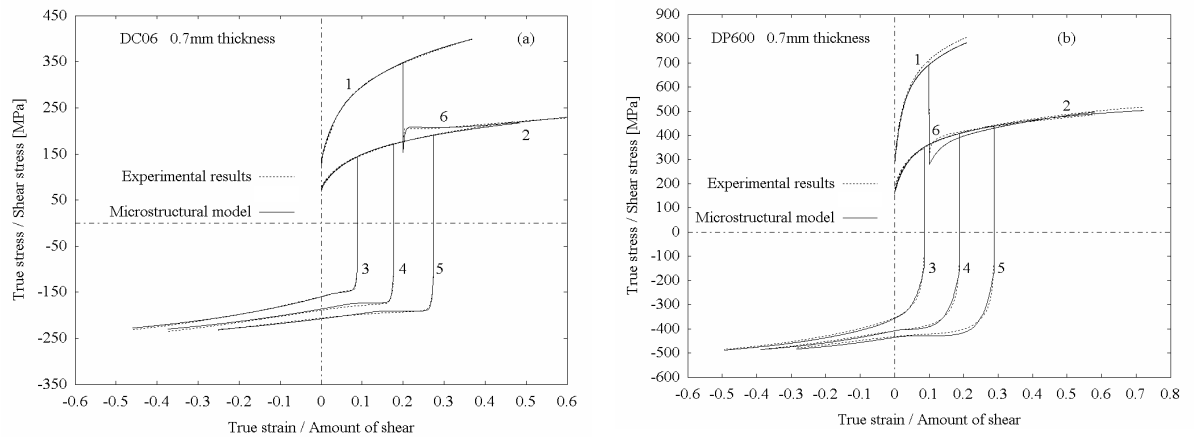


Figure 2. Comparison of mechanical tests with the prediction of the dislocation-based model in its (a) complete form for the IF steel DC06, (b) simplified form for the dual phase steel DP600. (1) Uniaxial tensile test. (2) Monotonic simple shear test. (3), (4), (5) Bauschinger simple shear test after 10%, 20% and 30% amount of shear in the forward direction. (6) Orthogonal test : simple shear in the rolling direction following a 20% true tensile strain in the same direction (after [11]).

2.6 Elastoplastic constitutive equations

As already mentioned in Sect. 2, we assume that all objective time derivatives are calculated with the same corotational spin $\dot{\mathbf{R}}\mathbf{R}^{\mathrm{T}} = \mathbf{W}$. It then proves convenient to reformulate the constitutive and evolution equations in terms of 'rotation-compensated' quantities, which will be denoted by a superposed hat. More precisely, if $\mathbf{T}$ and $\mathbf{S}$ denote as before a second-order and a fourth-order tensor, respectively, then the corresponding rotation-compensated tensors, $\hat{\mathbf{T}}$ and $\hat{\mathbf{S}}$, will be defined by

$$T_{ij} = R_{ip}\, R_{jq}\, \hat{T}_{pq}, \qquad S_{ijkm} = R_{ip}\, R_{jq}\, R_{kr}\, R_{ms}\, \hat{S}_{pqrs}. \tag{2.39}$$

The main advantage of this transformation is that the Jaumann-type derivatives of the initial tensors are related to the material time derivatives of the rotation-compensated tensors by relations similar to (3.1), namely

$$\overset{\circ}{T}_{ij} = R_{ip}\, R_{jq}\, \dot{\hat{T}}_{pq}, \qquad \overset{\circ}{S}_{ijkm} = R_{ip}\, R_{jq}\, R_{kr}\, R_{ms}\, \dot{\hat{S}}_{pqrs}. \tag{2.40}$$

It is also noteworthy that the transformation (2.39) preserves the norms and the double-contracted tensor products.

With this notation, the main equations of the dislocation-based microstructural model presented in Sect. 2.5 may be rewritten as follows.

$$\Phi = \bar{\sigma} - Y = 0, \qquad Y = Y_0 + R + f\left\|\hat{\mathbf{S}}\right\|, \tag{2.41}$$

$$\bar{\sigma}^2 = \hat{\mathbf{s}} : \hat{\mathbf{M}} : \hat{\mathbf{s}}, \qquad \hat{\mathbf{s}} = \hat{\boldsymbol{\sigma}}' - \hat{\mathbf{X}}, \tag{2.42}$$

$$\hat{\mathbf{D}}^{\mathrm{p}} = \dot{\lambda}\,\hat{\mathbf{V}}, \qquad \hat{\mathbf{V}} = \frac{1}{\bar{\sigma}}\,\hat{\mathbf{M}} : \hat{\mathbf{s}}\,, \qquad \overset{\circ}{\hat{\boldsymbol{\sigma}}} = \hat{\mathbf{c}} : (\hat{\mathbf{D}} - \hat{\mathbf{D}}^{\mathrm{p}}), \qquad \hat{\mathbf{A}} = \frac{\hat{\mathbf{D}}^{\mathrm{p}}}{\left\|\hat{\mathbf{D}}^{\mathrm{p}}\right\|}, \tag{2.43}$$

$$\dot{R} = C_{\mathrm{R}}\,(R_{\mathrm{sat}} - R)\,\dot{\lambda}, \quad \overset{\circ}{\hat{\mathbf{X}}} = C_{\mathrm{X}}\left[(X_{\mathrm{sat}}/\bar{\sigma})\,\hat{\mathbf{s}} - \hat{\mathbf{X}}\right]\dot{\lambda}, \quad \overset{\circ}{\hat{\mathbf{P}}} = C_{\mathrm{P}}\,(\hat{\mathbf{A}} - \hat{\mathbf{P}})\,\dot{\lambda}. \tag{2.44}$$

$$S_{\mathrm{D}} = \hat{\mathbf{A}} : \hat{\mathbf{S}} : \hat{\mathbf{A}}, \qquad \hat{\mathbf{S}}_{\mathrm{L}} = \hat{\mathbf{S}} - S_{\mathrm{D}}\,\hat{\mathbf{A}} \otimes \hat{\mathbf{A}}, \qquad Z = \left\|\hat{\mathbf{S}}_{\mathrm{L}}\right\|, \tag{2.45}$$

$$\dot{S}_{\mathrm{D}} = C_{\mathrm{SD}}\left[\left(S_{\mathrm{sat}} - S_{\mathrm{D}}\right)g - S_{\mathrm{D}}\,h\right]\dot{\lambda}, \qquad \overset{\circ}{\hat{\mathbf{S}}}_{\mathrm{L}} = -C_{\mathrm{SL}}\left(Z/S_{\mathrm{sat}}\right)^{\mathrm{n}}\,\hat{\mathbf{S}}_{\mathrm{L}}\,\dot{\lambda}, \tag{2.46}$$

The plastic multiplier $\dot{\lambda}$ can be determined by imposing the consistency condition for plastic loading. The result depends on whether the time-marching scheme is implicit and involves consistent elastoplastic moduli, or it is explicit, and hence requires only the calculation of tangent elastoplastic moduli. For simplicity, we restrict here ourselves to consider the latter case, and hence to impose the consistency condition in the rate form. For the plastic loading, the yield condition $(2.41)_1$ has to be identically satisfied in a neighbourhood of time t and hence its time derivative at time t should vanish. By taking into account that

$$\|\,\hat{\mathbf{S}}\,\| = \sqrt{\hat{\mathbf{S}} : \hat{\mathbf{S}}} = \sqrt{S_{\mathrm{D}}^2 + Z^2},$$

this condition gives

$$\dot{\Phi} = \dot{\bar{\sigma}} - \dot{R} - \frac{S_{\mathrm{D}}\,\dot{S}_{\mathrm{D}} + Z\,\dot{Z}}{\sqrt{S_{\mathrm{D}}^2 + Z^2}} = 0. \tag{2.47}$$

On the other hand, we have

$$\begin{aligned}\dot{\bar{\sigma}} &= \frac{1}{\bar{\sigma}}\left(\hat{\mathbf{M}} : \hat{\mathbf{s}}\right) : \dot{\hat{\mathbf{s}}} = \hat{\mathbf{V}} : (\overset{\circ}{\hat{\boldsymbol{\sigma}}}' - \overset{\circ}{\hat{\mathbf{X}}}) = \hat{\mathbf{V}} : \hat{\mathbf{c}} : (\hat{\mathbf{D}} - \hat{\mathbf{D}}^{\mathrm{p}}) - \hat{\mathbf{V}} : \overset{\circ}{\hat{\mathbf{X}}} \\ &= \hat{\mathbf{V}} : \hat{\mathbf{c}} : \hat{\mathbf{D}} - \left[\hat{\mathbf{V}} : \hat{\mathbf{c}} : \hat{\mathbf{V}} + C_{\mathrm{X}}\left(X_{\mathrm{sat}} - \hat{\mathbf{V}} : \hat{\mathbf{X}}\right)\right]\dot{\lambda}\end{aligned}$$

$$S_{\mathrm{D}}\,\dot{S}_{\mathrm{D}} + Z\,\dot{Z} = \left\{C_{\mathrm{SD}}\left[\left(S_{\mathrm{sat}} - S_{\mathrm{D}}\right)g - S_{\mathrm{D}}\,h\right]S_{\mathrm{D}} - C_{\mathrm{SL}}\left(Z/S_{\mathrm{sat}}\right)^{\mathrm{n}}\,Z^2\right\}\dot{\lambda}.$$

Next, by substituting these last two expressions and $(2.44)_1$ into (2.47) and solving with respect to $\dot{\lambda}$, we obtain

$$\dot{\lambda} = \frac{\alpha}{f_0}\, \hat{\mathbf{V}} : \hat{\mathbf{c}} : \hat{\mathbf{D}}, \tag{2.48}$$

where $\alpha = 1$ for the plastic loading and $\alpha = 0$ for the neutral loading, the unloading or in the elastic state, while

$$\begin{aligned} f_0 \;&= \hat{\mathbf{V}} : \hat{\mathbf{c}} : \hat{\mathbf{V}} \;+\; C_{\mathrm{R}} \left(R_{\mathrm{sat}} - R \right) \;+\; C_{\mathrm{X}} \left(X_{\mathrm{sat}} \;-\; \hat{\mathbf{V}} : \hat{\mathbf{X}} \right) \\ &+\; \frac{f}{\sqrt{Z^2 + S_{\mathrm{D}}^2}} \left\{ C_{\mathrm{SD}} \left[\left(S_{\mathrm{sat}} - S_{\mathrm{D}} \right) g \;-\; S_{\mathrm{D}} h \right] S_{\mathrm{D}} - C_{\mathrm{SL}} \left(Z / S_{\mathrm{sat}} \right)^{\mathrm{n}} Z^2 \right\}. \end{aligned} \tag{2.49}$$

Finally, introducing (2.48) into $(2.43)_1$ and the result obtained into $(2.43)_3$ yields the tangent elastoplastic constitutive equations

$$\hat{\dot{\boldsymbol{\sigma}}} \;=\; \hat{\mathbf{c}}^{\mathrm{ep}} : \hat{\mathbf{D}}, \qquad \hat{\mathbf{c}}^{\mathrm{ep}} \;=\; \hat{\mathbf{c}} \;-\; \frac{\alpha}{f_0} \left(\hat{\mathbf{c}} : \hat{\mathbf{V}} \right) \otimes \left(\hat{\mathbf{c}} : \hat{\mathbf{V}} \right). \tag{2.50}$$

where $\hat{\mathbf{c}}^{\mathrm{ep}}$ are the tangent elastoplastic moduli.

When the elastic behaviour may be considered as isotropic, the tensor $\mathbf{c}$ of the second-order elastic constants has the components

$$c_{\mathrm{ijkl}} \;=\; \hat{c}_{\mathrm{ijkl}} \;=\; \lambda \delta_{\mathrm{ij}} \delta_{\mathrm{kl}} \;+\; \mu \left(\delta_{\mathrm{ik}} \delta_{\mathrm{jl}} + \delta_{\mathrm{il}} \delta_{\mathrm{jk}} \right), \tag{2.51}$$

where λ and μ are Lamé's constants. It may be easily verified that in this case, for any deviatoric and symmetric second-order tensor, e. g. $\mathbf{V}$, we have

$$\mathbf{c} : \mathbf{V} = \hat{\mathbf{c}} : \hat{\mathbf{V}} = 2\mu \hat{\mathbf{V}}. \tag{2.52}$$

By introducing the last relation into $(2.50)_2$, we deduce the corresponding expression of the tangent elastoplastic moduli:

$$\hat{\mathbf{c}}^{\mathrm{ep}} \;=\; \hat{\mathbf{c}} \;-\; \alpha \frac{4\mu^2}{f_0} \hat{\mathbf{V}} \otimes \hat{\mathbf{V}}. \tag{2.53}$$

As already mentioned, using the rotation-compensated quantities in the constitutive modelling leads to an easier finite element implementation, as the objective rates of tensor variables are replaced by usual time derivatives. Moreover, by adequately choosing the orientation of the rotating frame (e.g. the orthotropic frame of a rolled sheet), it is often possible to assume that both the elastic and plastic parameters intervening in Eqs. (2.50) remain constant at each material point throughout the motion. Notwithstanding, when considering the global equilibrium equations, it is necessary to use a common frame and thus to rewrite Eq. $(2.50)_1$ in the form

$$\overset{\mathrm{o}}{\boldsymbol{\sigma}} \;=\; \mathbf{c}^{\mathrm{ep}} : \mathbf{D}, \tag{2.54}$$

where

$$c_{ijkm}^{\mathrm{ep}} \;=\; R_{pi}\, R_{qj}\, R_{rk}\, R_{sm}\, \hat{c}_{pqrs}^{\mathrm{ep}}. \tag{2.55}$$

In particular, as will be shown in the next section, Eq. (2.54) can be directly used for calculating the tangent stiffness matrix occurring in the principle of virtual power.

3. FINITE ELEMENT IMPLEMENTATION

We will illustrate the finite element implementation of the constitutive modelling developed in the preceding section on the case of sheet metal forming. A large variety of FE formulations have been developed in this area, but no single software can reliably simulate all types of forming processes and predict the eventual occurrence of forming defects. We shall recall here briefly some of the merits and drawbacks of three main types of such FE approaches, namely the static explicit, static implicit and dynamic explicit algorithms (for a more detailed comparison of these time-marching schemes, we refer to Makinouchi *et al.* [20,21]).

3.1 Principle of virtual power

Due to the incremental character of the elastic behaviour, it is convenient to adopt an updated Lagrangian description of the deformation process. Namely, the configuration of the sheet at time t is taken as reference configuration for the lapse of time between t and $t+\Delta t$, at the end of which the configuration of the sheet, the state variables, and the boundary conditions are updated. Then, the new configuration of the sheet is taken as reference configuration for the next time increment, and so on.

By time-differentiating the quasi-static equilibrium equations, in the absence of body forces, we obtain

$$\frac{\partial \dot{S}_{ij}}{\partial x_j} = 0, \qquad i,j = 1,2,3, \tag{3.1}$$

where $\mathbf{S}$ is the first Piola – Kirchhoff stress tensor referred to the configuration of the sheet at the beginning of the time increment (not to be confused, of course, with the fourth-order tensor $\mathbf{S}$ used in Sect. 2 to describe the directional strength of dislocation structures), and x_i are the Cartesian co-ordinates of a current point of the sheet at the beginning of the increment.

As regards the boundary conditions, we assume that the surface S of the sheet can be divided at time t into three parts: a part S_1, on which the rate of the nominal stress vector $\dot{\mathbf{s}}$ is prescribed, a second part S_2, on which the velocity vector $\mathbf{v}$ is prescribed, and a third part S_3, on which slipping conditions between the sheet and the tools, with Coulomb friction, are prescribed. Consequently, we may write

$$\dot{S}_{ij}\, n_j = \dot{s}_i^* \ \text{ on } S_1, \qquad v_i = v_i^* \ \text{ on } S_2, \tag{3.2}$$

where $\mathbf{n}$ is the unit outward normal to the sheet at time t, whereas $\dot{\mathbf{s}}^*$ and $\mathbf{v}^*$ are known functions of place and time. On the slipping surface S_3 we have

$$v_n^{\text{sheet}} = v_n^{\text{tool}}, \qquad \mathbf{t}_\tau = \mu_f \, | \, | \, v_\tau^{\text{rel}} / \| v_\tau^{\text{rel}} \|, \tag{3.3}$$

where v_n and t_n are the components of the velocity vector $\mathbf{v}$ and of the Cauchy stress vector $\mathbf{t}$ on the normal $\mathbf{n}$, respectively, $\mathbf{t}_\tau = \mathbf{t} - t_n \mathbf{n}$ is the tangential component of $\mathbf{t}$, μ_f is the friction coefficient, and $\mathbf{v}_\tau^{\text{rel}} = \mathbf{v}_\tau^{\text{sheet}} - \mathbf{v}_\tau^{\text{tool}}$ is the tangential relative velocity of the sheet with respect to the tool. It may be shown [22] that the friction forces can be iteratively taken into account, Eqs. (3.3) being replaced by

$$v_n^{\text{sheet}} = v_n^{\text{tool}}, \qquad \dot{\mathbf{s}}_\tau = \dot{\mathbf{s}}_\tau^* \ \text{ on } S_3, \tag{3.4}$$

where $\dot{\mathbf{s}}_\tau^*$ is a known function at each iteration.

In the present context, a velocity field $\mathbf{v}$ is said to be kinematically admissible if it satisfies Eqs. $(3.2)_2$ and $(3.3)_1$. A vector field $\delta\mathbf{v}$ is said to be a virtual velocity field if it satisfies the homogeneous boundary conditions

$$\delta\mathbf{v} = \mathbf{0} \ \text{ on } S_2, \qquad \delta v_n = 0 \ \text{ on } S_3. \tag{3.5}$$

With these definitions, it may be proved that the following theorem holds.

Principle of virtual power. A kinematically admissible velocity field $\mathbf{v}$ satisfies the boundary-value problem defined by Eqs. (3.1), (3.2) and (3.4) if and only if the condition

$$\int_V \dot{S}_{ij}\, \delta L_{ij}\, dV = \int_{S_1} \dot{s}_i^*\, \delta v_i\, dS + \int_{S_3} \dot{s}_{\tau i}^*\, \delta v_{\tau i}\, dS, \tag{3.6}$$

where $\delta L_{ij} = \partial(\delta v_i)/\partial x_j$, is satisfied for any virtual velocity field $\delta\mathbf{v}$. Here V and S denote, respectively, the region occupied by the sheet at time t and its boundary.

McMeeking and Rice [23] have proposed a slightly different form of this principle, which makes use directly of the Cauchy stress tensor $\boldsymbol{\sigma}$, namely

$$\int_V \{(\overset{\circ}{\tau}_{ij} - 2\sigma_{ik} D_{kj})\delta D_{ij} + \sigma_{jk} L_{ik} \delta L_{ij}\}\, dV = \int_{S_1} \dot{s}_i^*\, \delta v_i\, dS + \int_{S_3} \dot{s}_{\tau i}^*\, \delta v_{\tau i}\, dS. \tag{3.7}$$

Here $\boldsymbol{\tau} = (\det \mathbf{F}^e)\boldsymbol{\sigma}$ denotes the Kirchhoff stress and $\overset{\circ}{\boldsymbol{\tau}}$ its Jaumann derivative. Apparently more complicated than Eq. (3.6), the form (3.7) of the principle has nevertheless the advantage of involving tensors that occur directly in the formulation of the constitutive laws. Actually, for sheet metal forming, $\overset{\circ}{\boldsymbol{\tau}}$ can be replaced with a good approximation by $\overset{\circ}{\boldsymbol{\sigma}}$, which is directly provided by the tangent elastoplastic constitutive equations, as shown at the end of Sect. 2.6.

3.2 Static explicit algorithms

Clearly, the inviscid plastic behaviour described in Sect. 2, in particular the tangent elastoplastic constitutive equation (2.54), is insensitive to the choice of the time scale. Consequently, all time derivatives of the fields involved in the principle of virtual work (3.7) and in the constitutive and evolution equations can be replaced by the increments of these

fields corresponding to a trial increment $\mathrm{d}w$ of a monotonously increasing loading parameter w, e.g. the controlled displacement of a stamping tool.

Performing a finite element discretization, namely dividing the sheet into finite elements and assuming that the same shape functions are used to approximate the incremental displacement field $\mathrm{d}\mathbf{u}$ and the virtual displacement $\delta(\mathrm{d}\mathbf{u})$ corresponding to the trial increment $\mathrm{d}w$ of the loading parameter, we arrive in a standard way to a system of linear algebraic equations of the matrix form

$$\left[K_{\mathrm{T}}\right]\left\{\mathrm{d}U\right\}=\left\{\mathrm{d}F\right\}, \tag{3.8}$$

where $\left[K_{\mathrm{T}}\right]$ is the tangent stiffness matrix, while $\left\{\mathrm{d}U\right\}$ and $\left\{\mathrm{d}F\right\}$ denote, respectively, the arrays of trial incremental nodal displacements and forces corresponding to $\mathrm{d}w$.

The monitoring of the time increments in the static explicit formulation is most often realized by means of the so-called $r_{\min}$-strategy (see, e.g. Kawka and Makinouchi [24]). Namely, after calculating the solution of system (3.8), the trial incremental displacement $\left\{\mathrm{d}U\right\}$ is weighted by a coefficient r, whose admissible value, denoted by $r_{\min}$, is chosen in such a way that no significant changes occur in the stiffness matrix during the increment. Clearly, this condition can be satisfied only if the influence of the sources of non-linearities on the tangent approximation is limited by controlling the size of the increment.

Thus, the calculation performed for each increment comprises two stages. At the first stage, one solves the system (3.8) and determines the values of r for which:

(i) the material at a Gauss point passes from the elastic to the plastic state or viceversa;

(ii) a free node gets into contact with the tools or, conversely, a contact node gets free;

(iii) a sticking node becomes sliding;

(iv) the largest absolute value of the incremental principal strains attains a prescribed upper limit, say $\Delta\varepsilon_{\max}$;

(v) the Euclidian norm of the incremental rotation attains a prescribed upper limit, say $\Delta\omega_{\max}$;

(vi) a change in the deformation process or a required output is attained.

The minimum of all these values is denoted by $r_{\min}$ and defines the real size $\Delta w = r_{\min}\mathrm{d}w$ of the incremental loading parameter.

The second stage of the incremental step includes the validation of all changes for which the corresponding r-value does not exceed $r_{\min}$ by more than a small tolerance. Finally, the configuration of the sheet, the Cauchy stresses, as well as the orientation of the orthotropy frames and the hardening variables, are updated at all Gauss points in an explicit way, i.e. assuming that the rates of all quantities are constant over the time increment. The algorithm of the static explicit formulation is shown in Box 1.

START

- Input and check data
- **Repeat**
  - Impose a trial increment dw of the loading parameter
  - Calculate the (tangent) stiffness matrix K_{T}
  - Solve the system of equations $\left[K_{\mathrm{T}}\right]\left\{\mathrm{d}U\right\}=\left\{\mathrm{d}F\right\}$ for the incremental displacement $\left\{\mathrm{d}U\right\}$
  - Determine $r_{\min}$
  - Update sheet and tool configurations and state variables
  - Update contact/uncontact and sticking/slipping boundary conditions
  - Output results

  until *the end of the process*

END

Box 1. Algorithm of the static explicit time integration.

Despite the restrictions imposed on the size of the time step, the use of the tangent increments is inherently associated with some second-order deviations from local and global equilibrium. Whereas the accumulation of such incremental errors is generally harmless for the simulation of the deformation process, it can become critical e.g. for the prediction of the springback. Recently, this drawback has been largely eliminated by introducing in the static explicit formulation a new algorithm, called ALGONEQ, for systematically cancelling the non-equilibrated forces [3]. More precisely, the non-equilibrated nodal forces (and moments in the case of shell formulations) are calculated at the end of each increment. Then, whenever their maximum norm becomes higher than a certain tolerance multiplied by the maximum norm of the external forces, the non-equilibrated forces are cancelled by applying them with opposite signs on the sheet. An important feature of this algorithm is that this operation is generally performed in a single step, by using the corresponding tangent matrix, while the contact/sliding boundary conditions are corrected after updating the configuration of the sheet and the state variables ; thus, there are no iterations to perform and hence no convergence problems.

Notwithstanding, the static explicit algorithms are still suffering from the severe limitation of the time increments that is necessary in order to maintain the deviations from linearity within admissible tolerances. The implicit algorithms, to which we will turn now, are mainly intended to reduce this computational effort, for a given level of accuracy.

3.3 Static implicit algorithms

The main difference between the static implicit algorithms and the explicit ones is that the former employ a non-linear state-update algorithm over each increment and impose the consistency and equilibrium conditions on the final configuration of the increment. A large number of implicit algorithms have been proposed in the literature, which differ by the hypotheses made on the evolution of various non-linear geometric and physical quantities during each increment, by the way of taking into account this evolution in calculating the stiffness matrix and by the iterative methods employed to assure the equilibrium and the contact/friction conditions. Therefore, we will merely illustrate this class of algorithms on a particular example, called semi-implicit algorithm [25,26], which represents a good compromise between accuracy and computational effort and has been successfully used for simulating rather complex sheet metal forming processes [27,28].

The semi-implicit algorithm, which is based on a classical predictor – corrector scheme, is presented in Box 2. It should be mentioned that in this particular case the predictor is essentially based on the $r_{\min}$-strategy explained in the preceding section. However, in determining the value of $r_{\min}$, the condition (i) listed in Sect. 3.2 is suppressed, whereas conditions (iv) and (v) are significantly relaxed, thus reducing the number of increments that are required to simulate the process. This time-strategy is justified by the improved accuracy of the state-update algorithm.

The essential difference between the semi-implicit algorithm and the explicit one is the introduction at each increment of an equilibrium loop, which is shown schematically in Box 3. First, the incremental strains and rotations are computed using the midpoint rule of Hughes and Winget [29]. Then, the consistency condition is imposed by using a generalized midpoint rule (see, e.g. Pinski *et al.* [30]) and the resulting non-linear equation is solved at each integration point by means of a Newton - Raphson iteration. After calculating the stress increments, the residual nodal forces $\{\Delta R\}$ are calculated on the last determined configuration of the sheet and the system

$$[K]\{\Delta U\} = \{\Delta R\}, \tag{3.9}$$

is solved to obtained the correction $\{\Delta U\}$ of the incremental displacements. This procedure is repeated until the norm of the residual forces lies within preset limits. To assure a quadratic convergence, one should use for $[K]$ in (3.9) the consistent stiffness matrix, which can be calculated by linearizing the stress-update algorithm around the last determined configuration of the sheet. Alternatively, it is possible to replace $[K]$ by the tangent stiffness matrix $[K_{\mathrm{T}}]$, calculated at the beginning of each equilibrium iteration. This option, which leads to a slower convergence rate, proves to be sometimes more cost-efficient, especially when the time increments are not very large.

The treatment of the contact/friction conditions in the present algorithm deserves a special explanation. Whereas conditions (ii) and (iii) in the determination of $r_{\min}$ are maintained,

START

- Input and check data
- **Repeat**
  - Impose a trial increment dw of the loading parameter
  - Calculate the (tangent) stiffness matrix K_{T}
  - Solve the system of equations $\left[K_{\mathrm{T}}\right]\left\{\mathrm{d}U\right\}=\left\{\mathrm{d}F\right\}$ for the incremental displacement $\left\{\mathrm{d}U\right\}$
  - Determine $r_{\min}$
  - Update sheet and tool configurations and state variables
  - Update contact/uncontact boundary conditions
  - Equilibrium iterative loop (cf. Box 3)
  - Adjust boundary conditions
  - Output results

  until *the end of the process*

END

Box 2. Algorithm of the static semi-implicit time integration.

- **Repeat**
  - Calculate incremental strains and rotations by the mid-point rule
  - Integrate the constitutive laws by using the generalized midpoint rule and a Newton-Raphson algorithm
  - Calculate the residual forces $\left\{\Delta R\right\}$
  - Solve the system of equations $\left[K\right]\left\{\Delta U\right\}=\left\{\Delta R\right\}$ for the incremental displacement $\left\{\Delta U\right\}$
  - Determine $r_{\min}$
  - Update sheet configuration and state variables for fixed positions of the tools and a fixed value of the loading parameter

  until *the norm of the residual forces lies within preset limits*

Box 3. Algorithm of the equilibrium iterative loop occurring in the static semi-implicit time integration.

only the contact/uncontact changes are validated at the end of the predictor phase, the sticking/slipping changes being considered too sensitive to a correct evaluation of the nodal forces. Finally, as shown in Box 2, all contact and friction boundary conditions are simply "adjusted" at the end of the equilibrium loop. Although this adjustment introduces new non-equilibrated forces, they are not cancelled by a new equilibrium loop, before going to the next step. This option, which justifies the name 'semi-implicit' given to the algorithm, diminishes somewhat the accuracy of the simulation, but proves to significantly increase its robustness.

The evolution towards fully implicit algorithms may be done in two different ways. A direct generalization of the preceding algorithm is the introduction of a contact/friction loop that incorporates the equilibrium one. More precisely, at the end of each equilibrium loop, the contact/friction conditions are updated. This generates, as already mentioned, some new non-equilibrated forces, which may be cancelled by a new equilibrium loop, and so on, until no changes occur in the contact/friction conditions. However, for problems involving a large number of sheet nodes, this latter situation can hardly be attained, and the iterations have to be limited by arbitrarily choosing their maximum number.

A second possible option is to use an augmented Lagrangian approach of the contact/friction. This leads to a non-linear system of equations whose unknowns are both the nodal displacements and the friction forces of the contact nodes and which can be solved within a unique iteration loop. For a thorough analysis of various strategies that can be used in this context and for their application to the simulation of forming processes, we refer to recent publications by Menezes and coworkers (see [31-33], where further references on this topic can be found).

Clearly, the main advantage of the implicit algorithms is their accuracy, which may be essential, e.g. when predicting springback. On the other hand, the convergence of the iteration schemes used in such formulations is not automatically assured, except for rather simple cases.

3.4 Dynamic explicit algorithms

Dynamic explicit algorithms are very robust and efficient for large-scale problems. The central difference explicit scheme is used to integrate the equations of motion, whereas the non-equilibrated forces are transformed into inertial forces at each step. Lumped mass matrices are used, and hence no system of equations has to be solved.

Despite its success for industrial applications, dynamic explicit codes have also some intrinsic drawbacks. Thus, in order to reduce the number of steps necessary to simulate the almost quasi-static forming processes, several numerical artefacts have to be employed, e.g. the increase of the mass density and of the punch velocity by at least one order of magnitude and the introduction of an artificial damping, in order to limit the inertial effects. Moreover, the results obtained when simulating the springback, depend on the type and dimensions of the finite elements and even on the number of integration points (Matiasson et al. [34]). Thus, the simulation of forming defects requires a considerable experience on the user side for adequately designing the finite element mesh and choosing the scaling parameters for mass, velocity and damping (see, e.g. Lee and Yang [35]).

4. CONCLUSION

The constitutive models analysed in Sect. 2 are able to satisfactorily predict the complex behaviour of several steels and aluminium alloys used e.g. in the car manufacturing. They can be easily implemented in finite element codes and are not very time-consuming, as they imply uniquely calculations restricted to each Gauss point. On the other hand, all three classes of time-marching schemes presented in Sect. 3, namely the static explicit, static implicit and dynamic explicit algorithms, have some specific merits and drawbacks, and hence do not permit so far to make a unique choice, even when limiting the application area to the sheet metal forming.

References

1. Lian, J., Barlat, F., Baudelet, B. (1989), *Plastic behaviour and stretchability of sheet metals. Part II: Effect of yield surface shape on sheet forming limit,* Int. J. Plasticity, **5**, 131-147.
2. Zhang, Z.T., Lee, D. (1995), *Effect of process variables and material properties on the springback of 2D-draw bending parts* (SAE Paper 950692 in SP-1067, Society of Automotive Sheet Engineers, Warrendale, PA), pp. 11-18.
3. Yamamura, N., Kuwabara, T., Makinouchi, A., Teodosiu, C. (2001), *Springback simulation by the static explicit FEM code, using a new algorithm for canceling the non-equilibrated forces,* in Proc. 7th Int. Conf. on Numerical Methods in Industrial Forming Processes (NUMIFORM'2001), Toyohashi, Japon.
4. Hu, Z., Rauch, E.F., Teodosiu, C. (1992), *Work-hardening behaviour of mild steel under stress reversal at large strains,* Int. J. Plasticity, **8**, 839-856.
5. Thuillier, S., Rauch, E.F. (1994), *Development of microbands in mild steel during cross loading,* Acta Metall. Mater., **42**, 1973-1983.
6. Nesterova, E.V., Bacroix, B., Teodosiu, C. (2001), *Microstructure and texture evolution under strain-path changes in low-carbon IF steel,* Metall. Mater. Trans., **A32**, 2527-2538.
7. Teodosiu, C., Hu, Z. (1995), *Evolution of the intragranular microstructure at moderate and large strains: modelling and computational significance,* in Proc. 5th Int. Conf. on "Numerical Methods in Industrial Forming Processes" (NUMIFORM' 95), Ithaca, USA, Eds. S.F. Shen, P.R. Dawson, Balkema, Rotterdam, pp. 173-182.
8. Teodosiu, C. (1997), *Plasticity of Single Crystals and Crystalline Aggregates,* in " Large Plastic Deformation of Crystalline Aggregates ", Lecture Notes Int. Centre for Mech. Sci. (Udine, 1996), Ed. C. Teodosiu, Springer, Berlin, pp. 21-80.
9. Teodosiu, C., Hu, Z. (1998), *Microstructure in the continuum modelling of plastic anisotropy,* in Proc. 19th Risø Int. Symp. on Materials Science, Risø National Laboratory, Roskilde, Denmark, pp. 149-168.

10. Haddadi, H., Bouvier, S., Levée, P. (2001), *Identification of a microstructural model for steels subjected to large tensile and/or simple shear deformations,* J. Physique IV France, **11**, 329-337.
11. Banu, M., Bouvier, S., Halim, H., Maier, C., Tăbăcaru, V., Teodosiu, C. (2001), *Selection and identification of the elastoplastic models for the materials used in the benchmarks of research project "Digital Die Design System"* (Report of LPMTM – CNRS, University Paris 13, Villetaneuse, France).
12. Teodosiu, C. (1970), *A dynamic theory of dislocations and its applications to the elastic-plastic continuum*, in Proc. Int. on "Fundamental Aspects of Dislocation Theory", Eds. J. A. Simmons, R. deWit, R. Bullough, N.B.S. Spec. Publ. 317, Washington D.C., USA, vol. 2, pp. 837-876.
13. Rice, J.R. (1971), *Inelastic constitutive relations for solids: an internal-variable theory and its application to metal plasticity,* J. Mech. Phys. Solids, **19**, 433-455.
14. Mandel, J. (1972), *Plasticité classique et viscoplsticité,* Lecture Notes Int. Centre for Mech. Sci. (Udine, 1971), Springer, Berlin.
15. Mandel, J. (1982), *Définition d'un repère privilégié pour l'étude des transformations anélastiques du polycrstal,* J. Méc. Théor. Appl., **1**, 7-23.
16. Sarma, G.B., Dawson, P.R. (1996), *Texture predictions using a polycrystal plasticity model incorporating neighbor interactions,* Int. J. Plasticity, **12**, 1023-1054.
17. Balasubramanian, S., Anand, L. (1996), *Single crystal and polycrystal elasto-viscoplasticity: application to earing in cup drawing of f.c.c. materials,* Computational Mechanics, **17**, 209 - 225.
18. Peeters, B., Seefeldt, M., Teodosiu, C., Kalidindi, S.R., Van Houtte, P., Aernoudt, E. (2001), *Work-hardening/softening behaviour of b.c.c. polycrystals during changing strain paths. I. An integrated model based on substructure and texture evolution, and its prediction of the stress-strain behaviour of an IF steel during two-stage strain paths,* Acta Materialia, **49**, 1607-1619.
19. Lemaître, J., Chaboche, J.-L. (1985), *Mécanique des matériaux solides* (Dunod, Paris).
20. Makinouchi, A., Teodosiu, C., Nakagawa, T. (1998), *Advances in FEM simulation and its related technologies in sheet metal forming*, Annals of CIRP, **47**, 641-649.
21. Makinouchi, A., Teodosiu, C. (2001), *Numerical methods for prediction of geometrical defects in sheet metal forming,* in Proc. 1st M. I. T. Conf. on "Computational Fluid and Solid Mechanics" (M. I. T., Cambridge, Ma., USA).
22. Teodosiu, C., Cao, H.-L. (1988), Residual stresses after axisymmetric deep drawing, in Proc. 15th Biennial Congress I.D.D.R.G. on "Controlling Sheet Metal Forming Processes, Dearborn, Michigan, USA, Ed. North American Deep Drawing Research Group: ASM International, pp. 309-319.
23. McMeeking, R.M., Rice, J.R. (1975), *Modelling large deformation anisotropic plastic behaviour of mild steel sheet,* Int. J. Solids Struct., **11**, 601-616.
24. Kawka, M., Makinouchi, A. (1995), *Shell-element formulation in the static explicit FEM code for the simulation of sheet stamping,* J. Mater. Process. Technol., **50**, 105-115.

25. Cao, H.-L. (1990), *Modélisation mécanique et simulation numérique de l'emboutissage (application à la deformation plane et axisymétrique).* Ph. D. Thesis, Inst. National Polytechnique de Grenoble, France.
26. Teodosiu, C., Cao, H.-L., Ladreyt, T., Detraux, J.M. (1991), *Implicit versus explicit methods in the simulation of sheet metal forming,* in "FE simulation of 3D sheet metal forming processes in automotive industry", VDI Berichte Nr. 894, Zürich, pp. 601-627.
27. Cao, H.-L., Teodosiu, C. (1992), *Numerical simulation of drawbeads for axisymmetric deep-drawing*, in Proc. Int. Conf. NUMIFORM'92 on "Numerical Methods in Industrial Forming Processes", Balkema, Rotterdam, pp. 439-448.
28. Teodosiu, C., Daniel, D., Cao, H.-L., Duval, J.-L. (1995), *Modelling and simulation of the can-making process using solid finite elements*, J. Mater. Process. Technol., **50**, 133-143.
29. Hughes T.J.R., Winget, J. (1980), *Finite rotation effects in numerical integration of rate constitutive equations arising in large deformation analysis,* Int. J. Numer. Meth. Engng., **15**, 1862-1867.
30. Pinski, P.M., Ortiz, M., Pister, K.S. (1982), *Numerical integration of rate constitutive equations in finite deformation analysis,* Comp. Meth. Appl. Mech. Engng., **40**, 137-158.
31. Menezes L.F. (1994), *Modelação tridimensional e simulação numérica dos processos de enformação por deformação plástica, aplicação à estampagem de chapas metálicas,* Ph. D. Thesis, Coimbra, Portugal.
32. Menezes, L.F., Teodosiu, C. (1999), *Improvement of the frictional contact treatment in a single loop iteration algorithm specific to deep-drawing simulations,* in Proc. Int. Conf. NUMISHEET'99 on "Numerical Simulation of 3D Sheet Forming Processes", Eds. J.C. Gélin, P. Picart, Besançon, France, vol. 1, pp. 197-202.
33. Menezes, L.F., Teodosiu, C. (2000), *Three-dimensional numerical simulation of the deep-drawing process using solid finite elements,* J. Mater. Proc. Technol., **97**, 100-106.
34. Mattiasson, K., Thilderkvist, P., Strange, A., Samuelsson, A. (1995), *Simulation of springback in sheet metal forming,* in Proc. Int. Conf. NUMIFORM'95, Eds. S. Shen, P.R. Dawson, Balkema, Rotterdam, pp. 115-124.
35. Lee, S.W., Yang, D.Y. (1998), An assessment of numerical parameters influencing springback in explicit finite element analysis of sheet metal forming processes, J. Mater. Process. Technol., **80-81**, 60-67.

A Note on Non-Newtonian Modelling of Blood Flow in Small Arteries

NADIR ARADA [1] and ADÉLIA SEQUEIRA[2]

Abstract

Due to the complex rheological behavior of blood flow, it is not possible to develop and computationally evaluate appropriate continuum constitutive models describing in particular the shear thinning and stress relaxation properties of blood flow. In this note we address in particular the well-posedness of the equations of motion of a specific shear-thinning viscoelastic model for blood flow in small arteries.

Key words. Blood rheology, Oldroyd-B fluids, viscoelasticity, shear-dependent viscosity, shear-thinning.

1 A brief introduction to blood rheology

Blood is a multi-component mixture with complex rheological characteristics. It consists of multiple particles namely red blood cells - RBCs (or erythrocytes), white blood cells - WBCs (or leucocytes), platelets and other matter, suspended in an aqueous polymer solution, the plasma (Newtonian fluid), containing inorganic and organic salts, proteins and transported substances. The haematocrit (cell matter that consists primarily of RBCs) forms approximately 45% of the volume of normal human blood.

In large and medium vessels, blood is usually modelled as a Newtonian liquid. However, in smaller vessels, with diameters comparable with those of the cells, blood behaves as a shear-thinning fluid. In particular, at rest or at low shear rates, blood seems to have a high apparent viscosity (due to RBCs aggregation into clusters called *rouleaux*) while at high shear rates the cells become disaggregated and deform into an infinite variety of shapes without changing volume (deformability of RBCs), resulting in a reduction in the blood's viscosity. The deformed RBCs align with the flow field and tend to slide upon plasma layers formed in between. Attempts to recognize the shear-thinning nature of blood were initiated by Chien *et al.* [7], [8] in the 1960s. Empirical models like the power-law, Cross [9], Carreau [6] or W-S generalized Newtonian fluid models [32] have been obtained by fitting experimental data in one dimensional flows (see Fig.1–2). Recently, Vlastos *et al.* [31] proposed a modified Carreau equation to capture the shear dependence of blood viscosity. Also the belief that blood demonstrates a yield shear stress led to one of the simplest constitutive models for blood, the Casson's equation [26].

[1]Centro de Matemática e Aplicações, Instituto Superior Técnico, Av. Rovisco Pais, 1049-001, Lisboa, Portugal; supported by Fundação para a Ciência e a Tecnologia, SFRH/BPD/3506/2000 (anadir@math.ist.utl.pt)

[2]Centro de Matemática e Aplicações and Departamento de Matemática, Instituto Superior Técnico, Av. Rovisco Pais, 1049-001, Lisboa, Portugal (adelia.sequeira@math.ist.utl.pt)

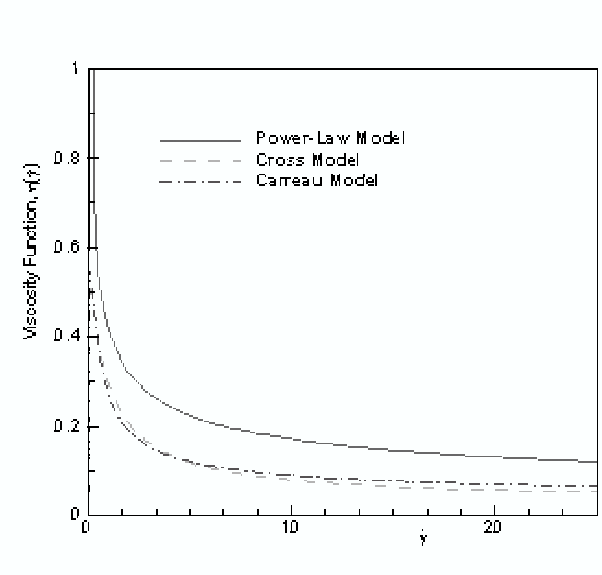


Figure 1: Blood viscosity as a function of shear rate for three generalized Newtonian fluid models.

However, none of these models are capable of describing the viscoelastic response of blood. Blood cells are essentially elastic membranes filled with a fluid and it seems reasonable, at least under certain flow conditions, to expect blood to behave like a viscoelastic fluid. At low shear rates RBCs aggregate and are 'solid-like', being able to store elastic energy that accounts for the memory effects in blood. Dissipation is primarily due to the evolution of the RBC networks and, given the paucity of data on temperature effects, the internal energy is assumed to depend only on the deformation gradient. At high shear rates, the RBCs disaggregate forming smaller rouleaux, and later individual cells, that are characterized by distinct relaxation times. RBCs become 'fluid-like', losing their ability to store elastic energy and the dissipation is primarily due to the internal friction. Upon cessation of shear, the entire rouleaux network is randomly arranged and may be assumed to be isotropic with respect to the current natural configuration. Thurston (see [27]) was among the earliest to recognize the viscoelastic nature of blood and that the viscoelastic behavior is less prominent with increasing shear rate. He proposed a generalized Maxwell model that was applicable to 1-D flow simulations ([28]) and observed later that, beyond a critical shear rate, the non-linear behavior is related to the microstructural changes that occur in blood ([29], [30]). Quemada [21] also proposed a non-linear Maxwell type model involving a first order kinetic equation used to determine a structural parameter related with the viscosity. Phillips and Deutsch [20] proposed a three dimensional frame invariant Oldroyd-B type model with four constants which could not capture the shear-thinning behavior of blood throughout the range of experimental data. The most recent three constant generalized Oldroyd-B model of Yeleswarapu *et al.* [35] is an improvement on the last model. It has been obtained by fitting experimental data in one dimensional flows and generalizing such curve fits to three dimensions. It captures the shear-thinning behavior of blood over a much larger range of shear rates but it has its limitations, given that the relaxation times do not depend on the shear rate, which does not agree with experimental observations.

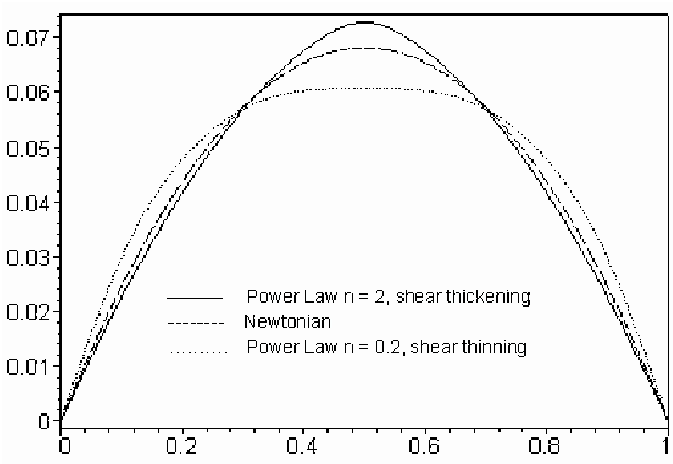


Figure 2: Velocity profiles for steady, fully developed flow in a straight pipe.

A general thermodynamic framework has been recently developed by Rajagopal and Srinivasa [23] for describing the response of bodies with multiple configurations. Rate type models due to Maxwell, Oldroyd and others which can also describe shear-thinning, can be generated within this framework. This approach is well suited for describing bodies whose response functions change with deformation and activation. More interestingly, it is possible to develop fluid models whose relaxation times depend on the shear rate and where, according to certain experimental observations, the viscoelastic character of blood becomes less important with increasing shear rate (see [2]).

While there has been a considerable research effort in blood rheology, the constitutive models have thus far focused on the aggregation and deformability of the RBCs, ignoring the role of platelets in the flow characteristics. Platelets are biconcave discoid cells containing various chemicals, much smaller than erytrocytes (approximately 6 μm^3 in size as compared to 90 μm^3) and forming a small fraction of the particulate matter of human blood (around 3% by volume). However they are by far the most sensitive of all the components of blood to chemical and physical agents, and play a significant role in blood rheology. Arterial occlusion, acute myocardial infarction, venous thrombosis and most strokes are some of the pathological processes related to platelet activation. Understanding these processes is an issue of major medical importance.

The mechanism of platelet activation and blood coagulation is quite complicated and not yet completely well understood. Recently, Kuharsky and Fogelson [15] have developed a model consisting of 59 first order ODEs that combines a fairly comprehensive description of coagulation biochemistry, interactions between platelets and coagulation proteins and effects of chemical and cellular transport. This model, as well as previous work developed along these lines (see *e.g.* [10], [33], [14]) can be considered as an important achievement to capture many of the biochemical aspects of the problem. However, they do not allow for the realistic hydrodynamical and rheological characteristics of blood flow in vessels whose geometry is made complex by the presence of wall-adherent platelets or atherosclerotic plaques. A phenomenological model recently introduced by Anand and Rajagopal [1] can be considered as the first approach to address this oversight.

2 Mathematical results for a shear-dependent viscoelastic model

The mathematical analysis and numerical simulation of the equations of motion of non-Newtonian viscoelastic fluids is a very challenging issue. The constitutive equations may lead to highly nonlinear systems of PDEs of a combined elliptic-hyperbolic type (or parabolic-hyperbolic, for unsteady flows) and the behavior of such equations is poorly understood. Special techniques of nonlinear analysis are needed to investigate questions of existence, uniqueness and stability of solutions and theoretical results are mainly based on 'small perturbations'. Usually the original nonlinear problem is written in a decoupled form, composed of a Stokes-like system and a scalar transport equation, that can be studied as two separate linear systems. The solvability of the original problem is established using a suitable fixed point argument. This technique has been successfully used in recent years for different viscoelastic fluids of differential and rate type, in several geometries (see *e.g.* [11] - [16], [17], [19], [24], or the monographs [34], [25] and the literature cited therein).

Numerical simulation is certainly considered as an important tool for prediction of non-Newtonian phenomena. In the last two decades, intensive research has been performed in this area, mainly for differential and rate-type models, using finite element or spectral methods for steady flows, finite differences in time and finite element or finite volume approximations in space for unsteady flows (see *e.g.* [13], the monograph [18] and references cited therein). The major drawback of many numerical schemes, due to the formidable amount of computation involved and to the loss of convergence for high values of the Weissenberg number (referred as the *'high Weissenberg number problem'*) is mainly related to the choice of improper boundary conditions and to the hyperbolic nature of the equations. One of the problems is that a straightforward Galerkin discretization of the constitutive law has poor stability properties if the advection term involving the velocity field and the stress tensor becomes dominant. The other problem is related to the mixed mathematical structure of the nonlinear systems whose behavior under discretization is poorly understood. Typically, specific numerical upwinding or artificial diffusivity techniques must be used together with appropriate choices of the spaces for velocities, stresses and pressure in such a way that the LBB inf-sup condition for velocity and pressure is satisfied and the stresses have higher accuracy than the velocities. In addition, advanced computational techniques such as highly adaptive refinement, parallel processing and novel matrix solvers will make the computations more affordable. The numerical schemes used for solving these complex systems of PDEs, in particular for the proposed blood flow models, must be based on a deep understanding of the mixed mathematical structure of the equations, in order to prevent numerical instabilities on problems that are mathematically well-posed.

As far as we know, generalizations of the above mentioned viscoelastic models incorporating a non-Newtonian viscosity function have not yet been studied from the mathematical point of view. The well-posedness of the equations of motion of a generalized Oldroyd-B fluid with shear-dependent viscosity, recently obtained by N. Arada and A. Sequeira [3] is a first step towards this aim. The model is able to capture shear-thinning and viscoelastic effects and can be considered thermodynamically based, in the simple case where the relaxation time is supposed to be a constant. The remaining of this note will be devoted to the proof of these mathematical results, following closely [3].

2.1 Formulation of the problem

We are concerned with flows of incompressible viscoelastic Oldroyd-B fluids with shear dependent viscosity in a bounded domain Ω of $I\!R^3$. For these fluids, the extra-stress tensor is related to the kinematic variables through

$$S + \lambda_1 \frac{\mathcal{D}S}{\mathcal{D}t} = 2\ (\nu + \nu_o(1 + |Dv|^2)^q)\ Dv + 2\ \lambda_2\ \frac{\mathcal{D}Dv}{\mathcal{D}t}, \tag{2.1}$$

where v is the velocity field, $Dv = \frac{1}{2}(\nabla v + \nabla v^t)$ denotes the symmetric part of the velocity gradient, $q \in]-\frac{1}{2}, 0[$, ν_o and ν are nonnegative real numbers satisfying $\nu + \nu_o > 0$, $\lambda_1 > 0$ and $\lambda_2 > 0$ are viscoelastic constants. The symbol $\frac{\mathcal{D}}{\mathcal{D}t}$ denotes the objective derivative of Oldroyd type defined by

$$\frac{\mathcal{D}S}{\mathcal{D}t} = \Big[\frac{\partial}{\partial t} + v \cdot \nabla\Big]\ S - S\ \nabla v^t - \nabla v\ S.$$

The Cauchy stress tensor is given by $T = -pI + S$, where p represents the pressure.
We decompose the extra-stress tensor S into the sum of its Newtonian part $\tau_s = 2\frac{\lambda_2}{\lambda_1}\ Dv$ and its viscoelastic part τ_e. It can be easily seen that the constitutive equation for τ_e is given by:

$$\tau_e + \lambda_1\ \frac{\mathcal{D}\tau_e}{\mathcal{D}t} = 2\Big(\nu + \nu_o(1 + |Du|^2)^q - \tfrac{\lambda_2}{\lambda_1}\Big)\ Dv.$$

Recalling the equations of conservation of momentum and mass in the domain Ω bounded in $I\!R^3$,

$$\rho\Big(\frac{\partial v}{\partial t} + v \cdot \nabla v\Big) = \nabla \cdot T + f, \qquad \nabla \cdot v = 0, \tag{2.2}$$

($\rho > 0$ is the constant mass density of the fluid, f denotes the external forces) and the conservation law given by (2.1), we look for steady solutions of the following system

$$\begin{cases} -\frac{\lambda_2}{\lambda_1}\ \Delta v + \rho\ v \cdot \nabla v + \nabla p = f + \nabla \cdot \tau_e & \text{in } \Omega, \\ \\ \nabla \cdot v = 0 & \text{in } \Omega, \\ \\ \tau_e + \lambda_1\ (v \cdot \nabla \tau_e - \tau_e \nabla v^t - \nabla v \tau_e) = 2\Big(\mu(|Dv|^2) - \frac{\lambda_2}{\lambda_1}\Big)\ Dv & \text{in } \Omega. \end{cases}$$

This system is supplemented by a Dirichlet homogeneous boundary condition

$$v = 0 \qquad \text{on } \partial\Omega. \tag{2.3}$$

We consider the dimensionless form of this system by introducing the following non-dimensional quantities

$$x = \frac{\tilde{x}}{L}, \quad v = \frac{\tilde{v}}{V}, \quad p = \frac{\tilde{p}L}{(\nu + \nu_o)V}, \quad \lambda_1 = \tilde{\lambda}_1, \quad \lambda_2 = \tilde{\lambda}_2,$$

where the symbol $\tilde{}$ is attached to dimentional parameters (V and L represent reference velocity and length). We also introduce the Weissenberg number $\mathcal{W}e = \frac{\lambda_1 V}{L}$ and the Reynolds number $\mathcal{R}e = \frac{\rho V L}{(\nu + \nu_o)}$. Finally, the dimensionless system takes the form

$$\begin{cases} -(1 - \varepsilon)\ \Delta v + \mathcal{R}e\ v \cdot \nabla v + \nabla p = f + \nabla \cdot \tau & \text{in } \Omega, \\ \\ \nabla \cdot v = 0 & \text{in } \Omega, \\ \\ v = 0 & \text{on } \partial\Omega. \end{cases} \tag{2.4}$$

$$\tau + \mathcal{W}e\ (v\cdot\nabla\tau + g(\tau,\nabla v)) = 2\Big(\tfrac{\nu_o}{\nu+\nu_o}((1+|Dv|^2)^q - 1) + \varepsilon\Big)Dv \qquad \text{in } \Omega, \tag{2.5}$$

with $g(\tau,\nabla v) = -\tau\ \nabla v^t - \nabla v\ \tau$ and $1-\varepsilon = \frac{\lambda_2}{\lambda_1\mu_o}$.

In all the sequel, we denote the norms in $H^k(\Omega)$ ($k \in I\!N$) by $\|\cdot\|_k$. We set

$$\mathcal{H} = \{v \in L^2(\Omega) \mid \nabla\cdot v = 0,\ v\cdot n = 0 \text{ on } \partial\Omega\},$$

where n is the unit outward normal vector to $\partial\Omega$. Endowed with the L^2-norm, $\mathcal{H}$ is a reflexive Banach space. Unless otherwise specified, C stands for a generic constant depending on Ω.

2.2 Formulation of an equivalent problem

Our goal in this section is to reduce the nonlinear system to an equivalent problem in a way that the ellipticity due to the viscoelastic terms becomes visible. This will lead to a reformulation of (2.4)-(2.5) as a fixed point equation.

We first recall some useful results concerning the transport equation, as well as properties and estimates of some operators. This is the aim of the following lemmas. The corresponding proofs can be found in [3] and [4].

Lemma 2.1 *Let $\Omega \subset I\!R^3$ be a bounded domain of class C^3, $\zeta \in \mathcal{H}$, $\delta \in]0,1[$ and $k \in \{0,1,2\}$. There exists a positive constant γ (depending on Ω) such that if*

$$2\gamma(1-\varepsilon,1)^+\mathcal{W}e\ \|\nabla\zeta\|_2 \le \delta, \tag{2.1}$$

then the following assertions are true

i) The operators $\mathcal{K}(\zeta) \equiv I + \mathcal{W}e\ \zeta\cdot\nabla$ and $\mathcal{K}_\varepsilon(\zeta) \equiv I + (1-\varepsilon)\ \mathcal{W}e\ \zeta\cdot\nabla$ are invertible with continuous inverse from their domain $\mathcal{D}_k(\zeta) = \{u \in H^k(\Omega) \mid \zeta\cdot\nabla u \in H^k(\Omega)\}$ into $H^k(\Omega)$.
ii) The operator $\mathcal{L}_\varepsilon(\zeta) \equiv (1-\varepsilon)I + \varepsilon\ \mathcal{K}(\zeta)^{-1}$ is an isomorphism in $H^k(\Omega)$ and

$$\mathcal{L}_\varepsilon(\zeta)^{-1} = \mathcal{K}_\varepsilon(\zeta)^{-1}\ \mathcal{K}(\zeta). \tag{2.2}$$

Moreover, the following estimates hold

$$\|\mathcal{K}(\zeta)^{-1}\|_k \le 2, \qquad \|\mathcal{K}_\varepsilon(\zeta)^{-1}\|_k \le 2,$$

$$\|\mathcal{L}_\varepsilon(\zeta)\|_k \le 2\ (1-\varepsilon+|\varepsilon|), \qquad \|\mathcal{L}_\varepsilon(\zeta)^{-1}\|_k \le 2\ \tfrac{1+|\varepsilon|}{1-\varepsilon}.$$

Lemma 2.2 *Let $\Omega \subset I\!R^3$ be a bounded domain of class C^3, ζ and $\widehat{\zeta} \in \mathcal{H}$ satisfying (2.1), $\delta \in]0,1[$ and $k \in \{0,1,2\}$. Then the following estimates hold*

$$\begin{aligned}
&\|\mathcal{K}(\zeta)^{-1}\phi - \mathcal{K}(\widehat{\zeta})^{-1}\widehat{\phi}\|_k \\
&\le C\Big(\|\phi-\widehat{\phi}\|_k + \mathcal{W}e\|\nabla(\zeta-\widehat{\zeta})\|_{(k,1)^+}\ (\|\phi\|_{k+1} + \|\widehat{\phi}\|_{k+1})\Big), \\
&\|\mathcal{K}_\varepsilon(\zeta)^{-1}\phi - \mathcal{K}_\varepsilon(\widehat{\zeta})^{-1}\widehat{\phi}\|_k \\
&\le C\Big(\|\phi-\widehat{\phi}\|_k + (1-\varepsilon)\mathcal{W}e\|\nabla(\zeta-\widehat{\zeta})\|_{(k,1)^+}\ (\|\phi\|_{k+1} + \|\widehat{\phi}\|_{k+1})\Big), \\
&\|\mathcal{L}_\varepsilon(\zeta)^{-1}\phi - \mathcal{L}_\varepsilon(\widehat{\zeta})^{-1}\widehat{\phi}\|_k \\
&\le C\Big(\tfrac{1+|\varepsilon|}{1-\varepsilon}\|\phi-\widehat{\phi}\|_k + |\varepsilon|\mathcal{W}e\|\nabla(\zeta-\widehat{\zeta})\|_{(k,1)^+}\ (\|\phi\|_{k+1} + \|\widehat{\phi}\|_{k+1})\Big),
\end{aligned}$$

for all $(\phi,\widehat{\phi}) \in (H^{k+1}(\Omega))^2$, where $C \equiv C(k,\Omega)$.

Let us now consider the system (2.4)-(2.5). By computing the divergence of both sides of equation (2.5), we obtain

$$\begin{aligned}
&\nabla \cdot \Big(\tau + \mathcal{W}e\ v \cdot \nabla\tau + \mathcal{W}e\ g(v,\tau)\Big) \\
&= \nabla \cdot \tau + \mathcal{W}e\ \Big(\partial v : \partial\tau + (v \cdot \nabla)\ \nabla \cdot \tau\Big) + \mathcal{W}e\ \nabla \cdot g(v,\tau) \\
&= 2\ \nabla \cdot \Big(\Big(\tfrac{\nu_o}{\nu+\nu_o}\Big((1+|Dv|^2)^q - 1\Big) + \varepsilon\Big) Dv\Big) \\
&= \varepsilon\Delta v + \tfrac{2\nu_o}{\nu+\nu_o}\ \nabla \cdot \Big(\Big((1+|Dv|^2)^q - 1\Big)\ Dv\Big),
\end{aligned}$$

where $(\partial v : \partial\tau)_i = \sum_{j,k} \frac{\partial u_k}{\partial x_j}\ \frac{\partial \tau_{ij}}{\partial x_k}$. Therefore, equation (2.5) becomes

$$\begin{aligned}
&\Big(I + \mathcal{W}e\ v \cdot \nabla\Big)\ \nabla \cdot \tau \\
&= \varepsilon\Delta v + \tfrac{2\nu_o}{\nu+\nu_o}\ \nabla \cdot \Big(\Big((1+|Dv|^2)^q - 1\Big)\ Dv\Big) - \mathcal{W}e\ \Big(\nabla \cdot g(v,\tau) + \partial v : \partial\tau\Big) \\
&\equiv \varepsilon\Delta v + F(v,\tau).
\end{aligned}$$

Supposing that there v satisfies (2.1), we deduce from assertion i) in Lemma 2.1 that

$$\nabla \cdot \tau = \varepsilon\ \mathcal{K}(v)^{-1}\ \Delta v + \mathcal{K}(v)^{-1}\ F(v,\tau).$$

We replace $\nabla \cdot \tau$ in (2.4) by its expression and get

$$-\mathcal{L}_\varepsilon(v)\ \Delta v + \mathcal{R}e\ v \cdot \nabla v + \nabla p = f + \mathcal{K}(v)^{-1}\ F(v,\tau). \tag{2.3}$$

By applying the operator $\mathcal{L}_\varepsilon(v)^{-1}$ to (2.3) and taking into account (2.2), we obtain

$$\begin{aligned}
&-\Delta v + \mathcal{R}e\ \mathcal{L}_\varepsilon(v)^{-1}(v \cdot \nabla v) + \mathcal{L}_\varepsilon(v)^{-1}\nabla p \\
&= \mathcal{L}_\varepsilon(v)^{-1} f + \mathcal{L}_\varepsilon(v)^{-1}\mathcal{K}(v)^{-1}\ F(v,\tau) = \mathcal{L}_\varepsilon(v)^{-1} f + \mathcal{K}_\varepsilon(v)^{-1}\ F(v,\tau).
\end{aligned}$$

After calculating the commutator of $\mathcal{L}_\varepsilon(v)^{-1}$ and ∇

$$\nabla(\mathcal{L}_\varepsilon(v)^{-1}p) - \mathcal{L}_\varepsilon(v)^{-1}\nabla p = \varepsilon\ \mathcal{W}e\ \mathcal{K}_\varepsilon(v)^{-1}\ [(\nabla v)^t \cdot \nabla(\mathcal{K}^{-1}(v)\ \mathcal{L}_\varepsilon(v)^{-1}p)], \tag{2.4}$$

we finally transform (2.4)-(2.5) into the following equivalent system

$$\begin{cases}
-\Delta v + \nabla(\mathcal{L}_\varepsilon(v)^{-1}p) = \mathcal{L}_\varepsilon(v)^{-1}\mathcal{F}(v) + \mathcal{K}_\varepsilon(v)^{-1}\widetilde{\mathcal{F}}(v,p,\tau) & \text{in } \Omega, \\
\nabla \cdot v = 0 & \text{in } \Omega, \\
v = 0 & \text{in } \partial\Omega,
\end{cases} \tag{2.5}$$

$$\tau + \mathcal{W}e\ v \cdot \nabla\tau = 2\varepsilon\ Dv + \mathcal{G}(v,\tau) \qquad \text{in } \Omega, \tag{2.6}$$

where

$$\begin{aligned}
&\mathcal{F}(v) = f - \mathcal{R}e\ v \cdot \nabla v, \\
&\widetilde{\mathcal{F}}(v,p,\tau) = \tfrac{2\nu_o}{\nu+\nu_o}\ \nabla \cdot \Big(\Big((1+|Dv|^2)^q - 1\Big)\ Dv\Big)
\end{aligned}$$

$$
\begin{aligned}
&-\mathcal{W}e\ \Big(\nabla\cdot g(v,\tau)+\partial v:\partial\tau\Big)\\
&+\varepsilon\ \mathcal{W}e\ (\nabla v)^t\cdot\nabla(\mathcal{K}(v)^{-1}\mathcal{L}_\varepsilon(v)^{-1}p),\\
\mathcal{G}(v,\tau)&=\tfrac{2\nu_o}{\nu+\nu_o}\Big(\Big((1+|Dv|^2)^q-1\Big)\ Dv\Big)-\mathcal{W}e\ g(v,\tau).
\end{aligned}
$$

The proof of existence and uniqueness of solutions to system (2.5)-(2.6) is based on the Banach fixed point theorem. More precisely, we define the mapping

$$\Phi:(\zeta,\pi,\vartheta)\longrightarrow(v,p,\tau),$$

through the Stokes system

$$
\begin{cases}
-\Delta v+\nabla(\mathcal{L}_\varepsilon(\zeta)^{-1}p)=\mathcal{L}_\varepsilon(\zeta)^{-1}\mathcal{F}(\zeta)+\mathcal{K}_\varepsilon(\zeta)^{-1}\widetilde{\mathcal{F}}(\zeta,\pi,\vartheta) & \text{in } \Omega,\\
\nabla\cdot v=0 & \text{in } \Omega,\\
v=0 & \text{in } \partial\Omega,
\end{cases}
\tag{2.7}
$$

and the transport equation

$$\tau+\mathcal{W}e\ \zeta\cdot\nabla\tau=2\varepsilon\ Dv+\mathcal{G}(\zeta,\vartheta)\qquad\text{in }\Omega,\tag{2.8}$$

and we look for a solution of (2.5)-(2.6) as a fixed point for the mapping Φ. We shall prove the following main result.

Theorem 2.1 *Let $\Omega\subset\mathbb{R}^3$ be a bounded domain of class C^3 and $f\in H^1(\Omega)$. Then, there exists a constant $\kappa>0$ such that if $\|f\|_1\le\kappa$, then problem (2.4)-(2.5) admits a unique solution $(v,p,\tau)\in H^3(\Omega)\times H^2(\Omega)\times H^2(\Omega)$. Moreover, the following estimate holds*

$$\|\nabla v\|_2+\|p\|_2+\|\tau\|_2\le C^{\frac{(1-\varepsilon+|\varepsilon|)(1+|\varepsilon|)}{1-\varepsilon}}\ \|f\|_1,$$

where $C\equiv C(\Omega)$.

2.3 Proof of the main result

Let us first state some estimates for the nonlinear terms that appear in our equivalent problem (2.5)-(2.6). The proof of this result can be found in [4] and is omitted here.

Lemma 2.3 *Let $\Omega\subset\mathbb{R}^3$ be a bounded domain of class C^3, ζ and $\widehat{\zeta}\in\mathcal{H}$ satisfying (2.1), $(\pi,\widehat{\pi},\vartheta,\widehat{\vartheta})\in(H^2(\Omega))^4$. Then the following estimates hold*

$$
\begin{aligned}
&\Big\|((1+|D\zeta|^2)^q-1)\ D\zeta\Big\|_2\le C\ \|\nabla\zeta\|_2^2\\
&\Big\|(\nabla\zeta)^t\cdot\nabla(\mathcal{K}(\zeta)^{-1}\mathcal{L}_\varepsilon(\zeta)^{-1}\pi)\Big\|_1\le C\ \tfrac{1+|\varepsilon|}{1-\varepsilon}\ \|\nabla\zeta\|_2\ \|\pi\|_2\\
&\Big\|((1+|D\zeta|^2)^q-1)\ D\zeta-((1+|D\widehat{\zeta}|^2)^q-1)\ D\widehat{\zeta}\Big\|_o\\
&\le C\ (\|\nabla\zeta\|_2+\|\nabla\widehat{\zeta}\|_2)\ (1+\|\nabla\zeta\|_2\|\nabla\widehat{\zeta}\|_2)\ \|\nabla(\zeta-\widehat{\zeta})\|_1,
\end{aligned}
$$

$$\Big\|(\nabla\zeta)^t\cdot\nabla(\mathcal{K}(\zeta)^{-1}\mathcal{L}_\varepsilon(\zeta)^{-1}\pi)-(\nabla\widehat{\zeta})^t\cdot\nabla(\mathcal{K}(\widehat{\zeta})^{-1}\mathcal{L}_\varepsilon(\widehat{\zeta})^{-1}\widehat{\pi})\Big\|_o$$

$$\leq C\Big(\tfrac{1+|\varepsilon|}{1-\varepsilon}\ \|\nabla(\zeta-\widehat{\zeta})\|_1\ \|\pi\|_1+\mathcal{W}e\ \|\nabla\widehat{\zeta}\|_1\tfrac{1+|\varepsilon|}{1-\varepsilon}\ \|\nabla(\zeta-\widehat{\zeta})\|_1\ (\|\pi\|_2+\|\widehat{\pi}\|_2)$$

$$+\|\nabla\widehat{\zeta}\|_1\Big[\tfrac{1+|\varepsilon|}{1-\varepsilon}\ \|\pi-\widehat{\pi}\|_1+|\varepsilon|\mathcal{W}e\ \|\nabla(\zeta-\widehat{\zeta})\|_1\ (\|\pi\|_2+\|\widehat{\pi}\|_2)\Big]\Big)$$

where $C\equiv C(\Omega)$.

For every $\delta\in]0,1[$, let $B(\delta)$ be the convex set defined by

$$B(\delta)=\Big\{(\zeta,\pi,\vartheta)\in\mathcal{H}\times H^2(\Omega)\times H^2(\Omega)\mid\|\nabla\zeta\|_2+\|\pi\|_2+\|\vartheta\|_2\leq\Lambda\delta\Big\}$$

where $\Lambda=\dfrac{1}{2\gamma(1-\varepsilon,1)^+\mathcal{W}e}$, and γ is the constant appearing in (2.1).

Proposition 2.1 *There exists $\kappa_o>0$ such that if $\|f\|_1\leq\kappa_o$, then Φ applies $B(\delta_o)$ into $B(\delta_o)$ for some $\delta_o\equiv\delta_o(\|f\|_1)<1$. Moreover, there exists $C\equiv C(\Omega)$ such that*

$$\Lambda\ \delta_o\leq C\ \tfrac{(1-\varepsilon+|\varepsilon|)(1+|\varepsilon|)}{1-\varepsilon}\|f\|_1.$$

Proof. Let $(\zeta,\pi,\vartheta)\in B(\delta)$. Classical results ensure existence of a unique solution $(v,p)\in\mathcal{H}\times H^2(\Omega)$ to the Stokes system (2.7). This solution satisfies

$$\|\nabla v\|_2+\|\mathcal{L}_\varepsilon(\zeta)^{-1}p\|_2\leq C\Big(\|\mathcal{L}_\varepsilon(\zeta)^{-1}\ \mathcal{F}(\zeta)\|_1+\|\mathcal{K}_\varepsilon(\zeta)^{-1}\ \widetilde{\mathcal{F}}(\zeta,\pi,\vartheta)\|_1\Big)$$

$$\leq C\ \Big(\|\mathcal{L}_\varepsilon(\zeta)^{-1}\|_1\ \|\mathcal{F}(\zeta)\|_1+\|\mathcal{K}_\varepsilon(\zeta)^{-1}\|_1\ \|\widetilde{\mathcal{F}}(\zeta,\pi,\vartheta)\|_1\Big),\tag{2.1}$$

where $C\equiv C(\Omega)$. Using the estimates stated in Lemma 2.1, it follows that

$$\|\nabla v\|_2+\|p\|_2\leq C\Big(1+\|\mathcal{L}_\varepsilon(\zeta)\|_2\Big)\Big(\|\nabla v\|_2+\|\mathcal{L}_\varepsilon(\zeta)^{-1}p\|_2\Big)$$

$$\leq C\Big(1+\|\mathcal{L}_\varepsilon(\zeta)\|_2\Big)\Big(\|\mathcal{L}_\varepsilon(\zeta)^{-1}\|_1\ \|\mathcal{F}(\zeta)\|_1+\|\mathcal{K}_\varepsilon(\zeta)^{-1}\|_1\ \|\widetilde{\mathcal{F}}(\zeta,\pi,\vartheta)\|_1\Big)$$

$$\leq C(1-\varepsilon+|\varepsilon|))\ \Big(\tfrac{1+|\varepsilon|}{1-\varepsilon}\ \|\mathcal{F}(\zeta)\|_1+\|\widetilde{\mathcal{F}}(\zeta,\pi,\vartheta)\|_1\Big).\tag{2.2}$$

On the other hand, in view of Assertion i) in Lemma 2.1, equation (2.8) admits a unique solution $\tau=\mathcal{K}^{-1}(\zeta)(2\varepsilon\ Dv+\mathcal{G}(\zeta,\vartheta))\in H^2(\Omega)$ satisfying

$$\|\tau\|_2\leq\|\mathcal{K}^{-1}(\zeta)\|_2\ \Big(|\varepsilon|\ \|\nabla v\|_2+\|\mathcal{G}(\zeta,\vartheta)\|_2\Big)\leq 2(1-\varepsilon+|\varepsilon|)\Big(\|\nabla v\|_2+\mathcal{G}(\zeta,\vartheta)\|_2\Big),$$

which together with (2.1) gives the estimate

$$\|\tau\|_2\leq C\ (1-\varepsilon+|\varepsilon|)\ \Big(\tfrac{1+|\varepsilon|}{1-\varepsilon}\|\mathcal{F}(\zeta)\|_1+\|\widetilde{\mathcal{F}}(\zeta,\pi,\vartheta)\|_1+\|\mathcal{G}(\zeta,\vartheta)\|_2\Big).\tag{2.3}$$

Combining (2.2) and (2.3), we obtain

$$\|\nabla v\|_2+\|p\|_2+\|\tau\|_2$$

$$\leq C\ (1-\varepsilon+|\varepsilon|)\ \Big(\tfrac{1+|\varepsilon|}{1-\varepsilon}\|\mathcal{F}(\zeta)\|_1+\|\widetilde{\mathcal{F}}(\zeta,\pi,\vartheta)\|_1+\|\mathcal{G}(\zeta,\vartheta)\|_2\Big).\tag{2.4}$$

Standard calculations give

$$\|g(\zeta,\vartheta)\|_2 + \|\partial\vartheta : \partial\zeta\|_1 \leq C\ \|\nabla\zeta\|_2\ \|\vartheta\|_2 \leq C\ (\Lambda\delta)^2, \tag{2.5}$$

$$\|\mathcal{F}(\zeta)\|_1 \leq C\ \Big(\|f\|_1 + \mathcal{R}e\ \|\nabla\zeta\|_2^2\Big) \leq C\ (\|f\|_1 + \mathcal{R}e\ (\Lambda\delta)^2). \tag{2.6}$$

Moreover, due to Lemma 2.3

$$\Big\|((1+|D\zeta|^2)^q - 1)\ D\zeta\Big\|_2 \leq C\ (\Lambda\delta)^2. \tag{2.7}$$

$$\Big\|(\nabla\zeta)^t \cdot \nabla(\mathcal{K}(\zeta)^{-1}\mathcal{L}_\varepsilon(\zeta)^{-1}\pi)\Big\|_1 \leq C\ \tfrac{1+|\varepsilon|}{1-\varepsilon}\ (\Lambda\delta)^2, \tag{2.8}$$

Taking into account (2.4)-(2.8) and the definition of $\mathcal{G}$ and $\widetilde{\mathcal{F}}$, we deduce that

$$\|\nabla v\|_2 + \|p\|_2 + \|\tau\|_2 \leq C(\varepsilon)\Big(\|f\|_1 + \Lambda_1\ (\Lambda\delta)^2\Big),$$

where $C(\varepsilon) = C\frac{(1-\varepsilon+|\varepsilon|)(1+|\varepsilon|)}{1-\varepsilon}$ and $\Lambda_1 = \mathcal{R}e + \frac{\nu_o}{\nu+\nu_o} + \mathcal{W}e\ (1+|\varepsilon|)$. Hence, one has $\Phi(B(\delta)) \subset B(\delta)$ provided that

$$C(\varepsilon)\Big(\|f\|_1 + \Lambda_1\ (\Lambda\delta)^2\Big) \leq \Lambda\delta.$$

By classical arguments we can show that if $\|f\|_1 < \min(\frac{\Lambda}{C(\varepsilon)}, \frac{\Lambda^2}{4C(\varepsilon)^2\Lambda_1})$, then there exist $\delta_o \equiv \delta_o(\|f\|_1) < 1$ and $\delta_1 > \delta_o$ such that the last condition is satisfied for every $\delta \in [\delta_o, \min(1,\delta_1)[$. Moreover $\Lambda\delta_o \leq C(\varepsilon)\|f\|_1$. ∎.

Proposition 2.2 *Let δ and ε_o be as in Proposition* 2.1. *There exists $\delta_o > 0$ such that if $\|f\|_1 \leq \delta_o$, then the mapping $\Phi : B(\varepsilon_o) \longrightarrow B(\varepsilon_o)$ is a contraction in $H^2(\Omega)\times H^1(\Omega)\times H^1(\Omega)$.*

Proof. Let (ζ,π,ϑ) and $(\widehat{\zeta},\widehat{\pi},\widehat{\vartheta})$ be in $B(\delta_o)$ and let (v,p,τ) and $(\widehat{v},\widehat{p},\widehat{\tau})$ be their respective images by Φ. Then

$$\begin{cases} -\Delta(v-\widehat{v}) + \nabla(\mathcal{L}_\varepsilon(\zeta)^{-1}(p-\widehat{p})) = \mathcal{F}_1 & \text{in } \Omega,\\ \nabla\cdot(v-\widehat{v}) = 0 & \text{in } \Omega,\\ v-\widehat{v} = 0 & \text{on } \partial\Omega \end{cases}$$

$$(\tau-\widehat{\tau}) + \mathcal{W}e\ \zeta\cdot\nabla(\tau-\widehat{\tau}) = 2\varepsilon\ D(v-\widehat{v}) + \mathcal{G}_1 \qquad \text{in } \Omega,$$

where

$$\begin{aligned}\mathcal{F}_1 &= \nabla((\mathcal{L}_\varepsilon(\widehat{\zeta})^{-1} - \mathcal{L}_\varepsilon(\zeta)^{-1})\widehat{p}) + \mathcal{L}_\varepsilon(\zeta)^{-1}\ \mathcal{F}(\zeta) - \mathcal{L}_\varepsilon(\widehat{\zeta})^{-1}\ \mathcal{F}(\widehat{\zeta})\\ &\quad + \mathcal{K}_\varepsilon(\zeta)^{-1}\ \widetilde{\mathcal{F}}(\zeta,\pi,\vartheta) - \mathcal{K}_\varepsilon(\widehat{\zeta})^{-1}\ \widetilde{\mathcal{F}}(\widehat{\zeta},\widehat{\pi},\widehat{\vartheta}),\\ \mathcal{G}_1 &= \mathcal{G}(\zeta,\vartheta) - \mathcal{G}(\widehat{\zeta},\widehat{\vartheta}) + \mathcal{W}e\ (\widehat{\zeta}-\zeta)\cdot\nabla\widehat{\vartheta}.\end{aligned}$$

Arguments similar to those used in the proof of Proposition 2.1 show that the triplet $(v-\widehat{v}, p-\widehat{p}, \tau-\widehat{\tau})$ satisfies

$$\begin{aligned}&\tfrac{1}{1-\varepsilon+|\varepsilon|}\Big(\|\nabla(v-\widehat{v})\|_1 + \|p-\widehat{p}\|_1 + \|\tau-\widehat{\tau}\|_1\Big) \leq C\ \Big(\|\mathcal{F}_1\|_o + \|\mathcal{G}_1\|_1\Big)\\ &\leq C\ \Big(\Big\|(\mathcal{L}_\varepsilon(\widehat{\zeta})^{-1} - \mathcal{L}_\varepsilon(\zeta)^{-1})\widehat{p}\Big\|_1 + \Big\|\mathcal{L}_\varepsilon(\zeta)^{-1}\ \mathcal{F}(\zeta) - \mathcal{L}_\varepsilon(\widehat{\zeta})^{-1}\ \mathcal{F}(\widehat{\zeta})\Big\|_o\end{aligned}$$

$$
\begin{aligned}
&+\Big\|\mathcal{K}_\varepsilon(\zeta)^{-1}\ \widetilde{\mathcal{F}}(\zeta,\pi,\vartheta)-\mathcal{K}_\varepsilon(\widehat{\zeta})^{-1}\ \widetilde{\mathcal{F}}(\widehat{\zeta},\widehat{\pi},\widehat{\vartheta})\Big\|_o\\
&+\Big\|\mathcal{G}(\zeta,\vartheta)-\mathcal{G}(\widehat{\zeta},\widehat{\vartheta})\Big\|_1+\mathcal{W}e\|\nabla(\zeta-\widehat{\zeta})\|_1\ \|\widehat{\vartheta}\|_2\Big)\\
&\le C\Big(\tfrac{1+|\varepsilon|}{1-\varepsilon}\|\mathcal{F}(\zeta)-\mathcal{F}(\widehat{\zeta})\|_o+\|\widetilde{\mathcal{F}}(\zeta,\pi,\vartheta)-\widetilde{\mathcal{F}}(\widehat{\zeta},\widehat{\pi},\widehat{\vartheta})\|_o+\|\mathcal{G}(\zeta,\vartheta)-\mathcal{G}(\widehat{\zeta},\widehat{\vartheta})\|_1\\
&+\mathcal{W}e\Big(\|\widehat{\vartheta}\|_2+|\varepsilon|\ (\|\widehat{p}\|_2+\|\mathcal{F}(\zeta)\|_1+\|\mathcal{F}(\widehat{\zeta})\|_1)\\
&+(1-\varepsilon)\ (\|\widetilde{\mathcal{F}}(\widehat{\zeta},\widehat{\pi},\widehat{\vartheta})\|_1+\|\widetilde{\mathcal{F}}(\widehat{\zeta},\widehat{\pi},\widehat{\vartheta})\|_1)\Big)\|\nabla(\zeta-\widehat{\zeta})\|_1\Big)\\
&\le C\Big\{\tfrac{1+|\varepsilon|}{1-\varepsilon}\|\mathcal{F}(\zeta)-\mathcal{F}(\widehat{\zeta})\|_o+\|\widetilde{\mathcal{F}}(\zeta,\pi,\vartheta)-\widetilde{\mathcal{F}}(\widehat{\zeta},\widehat{\pi},\widehat{\vartheta})\|_o+\|\mathcal{G}(\zeta,\vartheta)-\mathcal{G}(\widehat{\zeta},\widehat{\vartheta})\|_1\Big\}\\
&+C\Big\{\mathcal{W}e\Big(\|\widehat{\vartheta}\|_2+|\varepsilon|\ (\|\widehat{p}\|_2+\|f\|_1+\mathcal{R}e(\|\nabla\zeta)\|_2^2+\|\nabla\widehat{\zeta})\|_2^2)\Big)\\
&+(1-\varepsilon)\mathcal{W}e\Big(\tfrac{\nu_o}{\nu+\nu_o}\ (\|\nabla\zeta\|_2^2+\|\nabla\widehat{\zeta}\|_2^2)+\mathcal{W}e\ (\|\vartheta\|_2\|\nabla\zeta\|_2+\|\widehat{\vartheta}\|_2\|\nabla\widehat{\zeta}\|_2)\Big)\\
&+|\varepsilon|(\mathcal{W}e)^2\tfrac{1+|\varepsilon|}{1-\varepsilon}\ \Big(\|\pi\|_2\|\nabla\zeta\|_2+\|\widehat{\pi}\|_2\|\nabla\widehat{\zeta}\|_2\Big)\Big\}\ \|\nabla(\zeta-\widehat{\zeta})\|_1. \qquad (2.9)
\end{aligned}
$$

On the other hand, classical calculations give

$$
\|\zeta\cdot\nabla\zeta-\widehat{\zeta}\cdot\nabla\widehat{\zeta}\|_o\le C\ (\|\nabla\zeta\|_2+\|\nabla\widehat{\zeta}\|_2)\ \|\nabla(\zeta-\widehat{\zeta})\|_1, \qquad (2.10)
$$

$$
\begin{aligned}
&\Big\|g(\zeta,\vartheta)-g(\widehat{\zeta},\widehat{\vartheta})\Big\|_1+\Big\|\partial\vartheta:\partial\zeta-\partial\widehat{\vartheta}:\partial\widehat{\zeta}\Big\|_o\\
&\le C(\|\nabla\widehat{\zeta}\|_2+\|\vartheta\|_2)\ \Big(\|\nabla(\zeta-\widehat{\zeta})\|_1+\|\vartheta-\widehat{\vartheta}\|_1\Big). \qquad (2.11)
\end{aligned}
$$

Taking into account (2.9)-(2.11), Lemma 2.3 and the definition of $\mathcal{F}$, $\widetilde{\mathcal{F}}$, $\mathcal{G}$, and with strightforward calculations, we finally obtain

$$
\begin{aligned}
&\|\nabla(v-\widehat{v})\|_1+\|p-\widehat{p}\|_1+\|\tau-\widehat{\tau}\|_1\\
&\le C^{\frac{(1+|\varepsilon|)(1-\varepsilon+|\varepsilon|)}{1-\varepsilon}}\ \Theta(\Lambda\delta_o)\ \|\nabla(\zeta-\widehat{\zeta})\|_1+\|\pi-\widehat{\pi}\|_1+\|\vartheta-\widehat{\vartheta}\|_1,
\end{aligned}
$$

where

$$
\begin{aligned}
\widehat{\Theta}(x)&=|\varepsilon|\mathcal{W}e\|f\|_1+(1+(1+|\varepsilon|)\mathcal{W}e)\ x+\mathcal{W}e\ (|\varepsilon|\mathcal{R}e+(1-\varepsilon+|\varepsilon|)\mathcal{W}e)\ x^2\\
&+\tfrac{\nu_o}{\nu+\nu_o}x\ (1+x(1-\varepsilon)\mathcal{W}e+x^2).
\end{aligned}
$$

The mapping Φ is then a contraction provided that

$$
C^{\frac{(1+|\varepsilon|)(1-\varepsilon+|\varepsilon|)}{1-\varepsilon}}\ \Theta(\Lambda\delta_o)<1.
$$

The proof is complete. ∎

The statement of Theorem 2.1 is a consequence of Proposition 2.1, Proposition 2.2 and the following version of the Banach fixed point theorem.

Theorem 2.2 *Let X and Y be Banach spaces such that X is reflexive and $X \hookrightarrow Y$. Let B be a non-empty, closed, convex and bounded subset of X and let $\Phi : B \longrightarrow B$ be a mapping such that*

$$\|\Phi(u) - \Phi(v)\|_Y \leq M \ \|u - v\| \qquad \text{for all } u, v \in B, \quad (0 \leq M < 1),$$

then Φ has a unique fixed point in B.

Aknowledgement. The authors would like to thank Professor K. R. Rajagopal for helpful discussions related to the viscoelastic fluid model involved in this research. This work has been partially supported by the FCT grant SFRH/BDP/3506/2000 (N. Arada), Center for Mathematics and its Applications (CEMAT) through FCT's Funding Program and by European Union FEDER/POCTI, Project POCTI/MAT/41898/ 2001.

References

[1] M. Anand, K. R. Rajagopal, A mathematical model to describe the change in the constitutive character of blood due to platelet activation, C.R. Mecanique, 330, 2002, 557-562.

[2] M. Anand, K. R. Rajagopal, A shear-thinning viscoelastic fluid model for describing the flow of blood, 2002 (submitted).

[3] N. Arada, A. Sequeira, Strong steady solutions for a generalized Oldroyd-B model with shear-dependent viscosity in a bounded domain, 2002 (accepted for publication in M3AS).

[4] N. Arada, A. Sequeira, Steady flow of a generalized Oldroyd-B fluid around an obstacle, 2003 (submitted)

[5] R. B. Bird, R. C. Armstrong, O. Hassager, Dynamics of Polymeric Liquids, 2nd edition, John Wiley & Sons, New York, 1987.

[6] P.J. Carreau, PhD Thesis, University of Wisconsin, Madison, 1968.

[7] S. Chien, S. Usami, R.J. Dellenback, M.I. Gregersen, Blood viscosity: Influence of erythrocyte deformation, Science 157 (3790), 1967, 827-829.

[8] S. Chien, S. Usami, R.J. Dellenback, M.I. Gregersen, Blood viscosity: Influence of erythrocyte aggregation, Science 157 (3790), 1967, 829-831.

[9] M. M. Cross, Rheology of non-Newtonian fluids: a new flow equation for pseudoplastic systems, J. Colloid Sci., 20, 1965, 417-437.

[10] A. L. Fogelson, Continuum models of platelet aggregation: formulation and mechanical properties, SIAM J. Appl. Math., 52, 1992, 1089.

[11] C. Guillopé, J. C. Saut, Existence results for the flow of viscoelastic fluids with a differential constitutive law, Nonlinear Anal., Th. Meth. & Appli., 15, 1990, 849-869.

[12] A. Hakim, Mathematical analysis of viscoelastic fluids of White-Metzner type, J. Math. Anal. Appl., 185, 1994, 675-705.

[13] R. Keunings, A survey of computational rheology, in: Proceedings of the XIIIth International Congress on Rheology (D.M. Binding *et al.* ed.), British Soc. Rheol., 1, 2000, 7-14.

[14] A. Kuharsky, Mathematical modeling of blood coagulation, PhD Thesis, Univ. of Utah, 1998.

[15] A. Kuharsky, A. L. Fogelson, Surface-mediaed control of blood coagulation: the role of binding site densities and platelet deposition, Biophys. J., 80 (3), 2001, 1050-1074.

[16] S.A. Nazarov, A. Sequeira, J. H. Videman, Asymptotic behaviour at infinity of three-dimensional steady viscoelastic flows, Pacific J. of Math., 203, 2, 2002, 461-488.

[17] A. Novotný, A. Sequeira, J. H. Videman, Steady motions of viscoelastic fluids in 3-D exterior domains - existence, uniqueness and asymptotic behavior, Arch. Rational Mech. Analysis, 149, 1999, 49-67.

[18] R. G. Owens, T. N. Phillips, Techniques of Computational Rheology, Imperial College Press/World Scientific, London, UK, 2002.

[19] K. Pileckas, A. Sequeira, J. H. Videman, Steady flows of viscoelastic fluids in domains with outlets to infinity, J. Math. Fluid Mech., 2, 2000, 185-218.

[20] W.M. Phillips, S. Deutsch, Towards a constitutive equation for blood, Biorheology, 12(6), 1975, 383-389.

[21] D. A. Quemada, A non-linear Maxwell model of biofluids - Application to normal blood, Biorheology, 30(3-4), 1993, 253-265.

[22] K. R. Rajagopal, Mechanics of non-Newtonian fluids, in: *Recent Developments in Theoretical Fluid Mechanics*,G.P. Galdi and J. Necas (eds),Pitman Research Notes in Mathematics Series, 291, Longman's Scientific and Technical, 1993, 129-162.

[23] K. R. Rajagopal, A. Srinivasa, A thermodynamic framework for rate type fluid models, J. of Non-Newtonian Fluid Mech., 88, 2000, 207-228.

[24] M. Renardy, Existence of slow steady flows of viscoelastic fluids with a differential constitutive equation, Z. Angew. Math. Mech., 65, 1985, 449-451.

[25] M. Renardy, Mathematical Analysis of Viscoelastic Flows, CBMS 73, SIAM, Philadelphia, 2000.

[26] G. W. Scott-Blair, An equation for the flow of blood, plasma and serum through glass capillaries, Nature, 183, 1959, 613 - 614.

[27] G. B. Thurston, Viscoelasticity of human blood, Biophys. J., 12, 1972, 1205-1217.

[28] G. B. Thurston, Rheological parameters for the viscosity, viscoelasticity and thixotropy of blood, Biorheology, 16, 1979, 149-162.

[29] G. B. Thurston, Light transmission through blood in oscillatory flow, Biorheology, 27, 1990, 685-700.

[30] G. B. Thurston, Non-Newtonian viscosity of human blood: flow-induced changes in microstructure, Biorheology, 31(2), 1994, 179-192.

[31] G. Vlastos, D. Lerche, B. Koch, The superimposition of steady and oscillatory shear and its effect on the viscoelasticity of human blood and a blood-like model fluid, Biorheology, 34(1), 1997, 19-36.

[32] F. J. Walburn, D. J. Schneck, A constitutive equation for whole human blood, Biorheology, 13, 1976, 201-210.

[33] N. T. Wang, A. L. Fogelson, Computational methods for continuum models of platelet aggregation, J. Comput. Phys., 151, 1999, 649-675.

[34] J. H. Videman, Mathematical analysis of viscoelastic non-Newtonian fluids, Phd Thesis, Instituto Superior Técnico, Lisbon, 1997.

[35] K. K. Yelesvarapu, M. V. Kameneva, K. R. Rajagopal, J. F. Antaki, The flow of blood in tubes: theory and experiment, Mech. Res. Comm, 25 (3), 1998, 257-262.

Advanced School and Workshop on Modelling and Numerical Simulation in Continuum Mechanics
Coimbra, 14-18 July 2003

FINITE ELEMENT SIMULATION OF SHEET METAL FORMING FROM AN INDUSTRIAL PERSPECTIVE

Kjell Mattiasson

Volvo Car Corporation
405 31 Göteborg, Sweden
e-mail: kmattias@volvocars.com

and
Department of Structural Mechanics
Chalmers University of Technology
412 96 Göteborg, Sweden
e-mail: kjell@sm.chalmers.se, web page: http://www.sm.chalmers.se/

Key words: Sheet metal, forming, simulation, Finite Element

Abstract. *During the last few years an enormous progress in the industrial use of numerical tools for the simulation of sheet forming processes has taken place. This is especially true for the automotive industry. The present paper tries to give a brief introduction to the subject of practical sheet metal forming, and to describe some of those forming defects, which can occur. Thereafter, a state-of-the-art review of methods and procedures for sheet forming simulation in practical use today will be presented. This concerns especially various Finite Element formulations used for the current application, including a brief historical review on the subject. Finally, some shortcomings of today's simulation technology will be described.*

1 INTRODUCTION

An often-cited statement is that sheet metal forming during the last decade has turned from being an art to being a science. The background to this statement is that sheet metal forming, from ancient to modern times, has been the task of skilled craftsmen rather than of theorists and scientists. Very few theoretical aids have been available for facilitating the die designers in their task, but they have had to rely on their own experience and simple guidelines. The design and tryout of forming tools have, thus, been a time consuming trial and error process.

However, the demand for shorter lead times, especially in the automotive industry, has accentuated the need for a computerized simulation aid, in which the forming process can be simulated, analyzed, and optimized, before any hard tools are built. During the last few years this desire has partially become reality, and today the simulation technique has been integrated in the die design process at many automotive and tool manufacturers.

The aim of the present paper is to give a state-of-the art review of current sheet forming simulation methods. The focus will be on the industrial implementation of these methods, rather than on current academic achievements. In order to provide a historical perspective on the subject, a brief review of the developments in this area during the last three decades will also be given.

2 SOME PRACTICAL ASPECTS ON SHEET METAL FORMING

2.1 Sheet forming processes

By far the most common sheet forming process is *stamping*, which especially is used in the huge automotive industry. In the stamping process the metal sheet is formed by rigid tools, which consist of a punch (male part), a die (female part), and, finally, a blankholder. The role of the blankholder is to press the blank against the die and prevent it from wrinkling, and also through friction forces control the material flow into the die cavity during the stroke. The main advantage of the stamping process is its high productivity, which is a very important quality in the highly efficient and automated car manufacturing industry. In Fig. 1 a Finite Element (FE) model of a stamping operation is displayed.

In *hydroforming* processes a hydraulic pressure replaces one of the rigid tools in the stamping process. For instance, in the *flex-forming* process the punch is replaced by a hydraulic pressure, which presses the blank down into the die cavity. Flex-forming is typically used in the aircraft industry, and for manufacturing of prototype parts in the automotive industry. The advantage of most hydroforming processes is that they allow parts to be manufactured, which would have been impossible in ordinary stamping. The main disadvantage is their low productivity, which makes them unsuitable to use in the automotive industry.

One type of hydroforming processes has, however, gained a great deal of popularity in the automotive industry in later years. This is the *tube hydroforming* process. In this process beam type parts with closed cross sections are manufactured. A tube-shaped work piece is first bent, and then formed to its final shape by an internal hydraulic pressure against an enclosing rigid

die. Although tube hydroforming is a rather slow process, it has gained a wide appreciation, since, due to its good formability, it allows a part to be manufactured in one piece, instead of being welded together of several stamped parts. A FE model of a tube hydroforming part can be seen in Fig. 2.

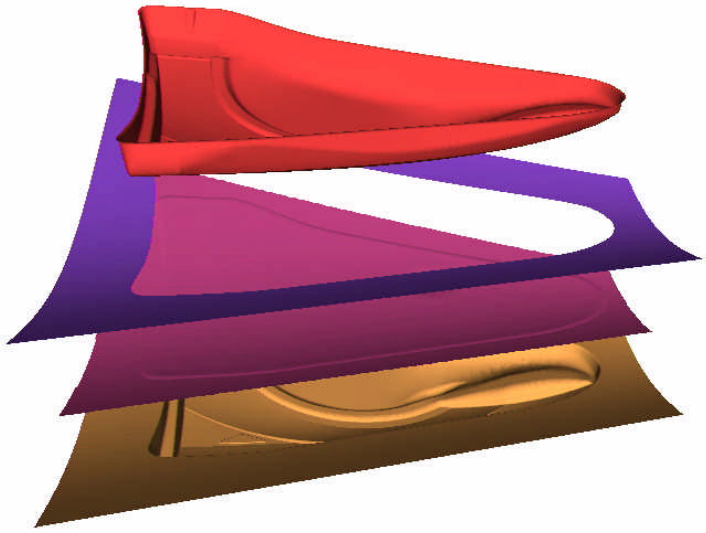


Figure 1 FE model of a stamping process. The parts are from top to bottom: punch, blankholder, blank, and die.

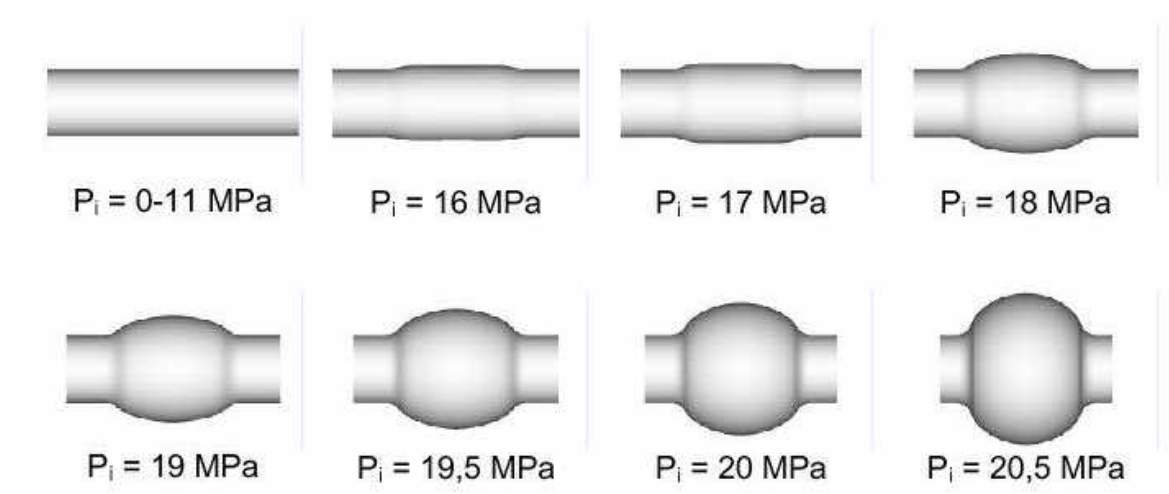


Figure 2 FE simulation of a tube hydroforming process

2.2 Forming defects

During the tryout of a forming process various types of defects are usually appearing in the formed part. Examples of such defects are:

- *Fracture* in the material, usually preceded by a marked strain localization.
- Excessive *thinning* in some areas of the blank
- *Wrinkling*, which implies the formation of bulges with relatively short wavelength due to high compressive stresses.
- *Buckling*, a term used for bulges with long wavelength, preferably appearing in unsupported areas of the blank with small compressive stresses.
- *Springback* is a term for those deformations that take place when a work piece is removed from the tools after completed forming.
- Various *surface defects*, which usually are due to insufficient stretching of the material.

It is the purpose of the die design and process layout work, and the subsequent tool tryout, to optimize the forming process in such a way that these defects can be avoided.

3 FINITE ELEMENT METHODS FOR SHEET FORMING SIMULATION

3.1 Introduction

The aim is that the simulation code should be able to simulate the complete forming process and to be able to unveil any possible defects. It should also be possible in the simulation to vary all those parameters, which in a practical case are used to optimize the process.

Sheet forming simulations tend to be very time consuming. One reason for this is that the process itself is computationally very complicated, involving effects such as nonlinear material behaviour, large deformations, and complicated contacts between tools and work piece. Another reason is that the FE models usually are very big, containing tens and hundreds of thousands elements. In the development of FE codes for sheet forming simulation, computational efficiency has therefore always been a primary concern.

3.2 Finite Element formulations

Through the years a number of different FE formulations for sheet forming simulation have been presented. These can differ from each other in several respects, such as FE types, kinematic description, constitutive description, and solution methodology. The bases for FE analysis of large deformation problems were not established until the mid 70's, and it was not until then the first procedures for FE simulation of sheet forming were presented.

The Hill'48 material model To be able to review the different FE formulations, we will first have a look at different ways of expressing the constitutive equations. The most commonly used constitutive relation in sheet metal forming contexts is the model of Hill[1] from 1948. It can describe orthonormal anisotropy of the material. The model is also known as Hill's quadratic model, since the stress terms describing the yield surface are all squared. The

effective stress can in matrix form be expressed as

$$\bar{\sigma} = \left(\{\sigma\}^T [A] \{\sigma\} \right)^{1/2} \tag{1}$$

where [A] is a matrix with constants describing the anisotropy of the material.

The normality condition can for associated plasticity be written

$$\{\dot{\varepsilon}^p\} = \dot{\lambda} \left\{ \frac{\partial f}{\partial \sigma} \right\} \tag{2}$$

In the special case of a quadratic yield condition this can, in view of Eq. (1), be expressed as

$$\{\dot{\varepsilon}^p\} = \frac{\dot{\lambda}}{\bar{\sigma}} [A] \{\sigma\} \tag{3}$$

Inverting this expression, and noting that $\dot{\lambda} = \dot{\bar{\varepsilon}}^p$, we get

$$\{\sigma\} = \frac{\bar{\sigma}}{\dot{\bar{\varepsilon}}^p} [A]^{-1} \{\dot{\varepsilon}^p\} \tag{4}$$

Note that this equation expresses *total* stress in terms of *rate* of plastic strain. Note also that it is only for quadratic yield conditions that the normality condition can be inverted to this form.

If the *plastic* strain rates in Eq. (4) are replaced by *total* strain rates, i.e. the elastic part of the deformation is ignored, this equation will form the basis of the *rigid-plastic* theory. A couple of the earlier FE formulations for sheet forming simulation were based on this form of the constitutive equations.

The flow formulation The *flow-formulation* for sheet metal forming is based on the above rigid-plastic material law. It uses a kind of Updated Eulerian formulation with nodal velocities as primary unknowns. The geometry is fixed in each time step, while the equilibrium is iteratively solved for. The geometry is then updated based on the calculated velocities.

It is interesting to note that there exists a complete analogy between the equations of the flow approach, and of those of small strain, linear elasticity. The only differences being that strain rate and nodal velocity in the flow formulation take the place of strain and nodal displacement in linear elasticity, and that the elasticity modulus in the elastic constitutive equations corresponds to a nonlinear ‘viscosity’ term in the rigid-plastic equations.

One of the main advantages of the flow approach is, thus, that the governing equations get a very simple appearance. A disadvantage of the approach is that problems occur when there are undeformed zones in the body, where $\dot{\bar{e}} = 0$, and the ‘viscosity’ turns to infinity. It takes some artificial actions to cure that problem. Another obvious disadvantage is of course that no phenomena related to elasticity, such as springback, can be simulated.

See for instance Onate et.al.[2] for further references on the subject.

The rigid-plastic formulation In the *rigid-plastic approach* the same rigid-plastic

constitutive relations as in the flow approach are used. However, some writers have preferred to rewrite these relations in terms of *increments* of strain. This leads naturally to a Langrangian FE formulation with nodal displacements as primary unknowns.

The disadvantages of such an approach are of course the same as for the flow approach. However, the present formulation do also lack the simplicity of the flow formulation, since the kinematic relations in a Lagrangian formulation are much more complicated than those in an Eulerian one.

Refs.[3,4,5] are examples of works in which the rigid-plastic approach has been employed.

The static-implicit method A sheet forming formulation, which is based on a Lagrangian description of motion and an elastic-plastic or elastic-viscoplastic constitutive law, is termed the *solid approach*. It is a formulation of considerable theoretical complexity, but has the advantage of being able to simulate also phenomena related to elasticity, such as springback. In contrast to the previous two approaches, this one is not restricted to quadratic yield conditions.

The resulting system of equations is normally solved by the Newton-Raphson method, or some similar technique. The method is in that case also known as the *static-implicit method.* In later years the importance of the concept of consistent linearisation has been realized. This concept is essential in order to preserve the quadratic rate of convergence of the Newton method, and has applicability both on stress integration as well as on contact/friction procedures. The use of consistent linearisation has implied a dramatic improvement of the performance of the solid approach, both with regard to efficiency as well as to robustness.

The main approximation introduced in the solid approach originates from the integration of the rate constitutive equations in order to calculate stresses.

The present approach has been used by numerous researchers, and the reader is referred to the proceedings from some of the recent conferences on the subject of metal forming simulation for further references. See for instance Sünkel et.al.[6], and Tang and Hu[7].

The static-explicit method The previous approaches have all been *implicit* in the sense that an iterative procedure has been employed in each step in order to fulfill the static equilibrium conditions. However, some authors have used a technique in which no iterations at all are performed. The updating of the geometry is just based on the result of the first iteration in each step. This implies that equilibrium is never satisfied. In order to reduce the errors involved, very small steps have to be taken. Several thousand steps are common for an ordinary simulation.

An advantage of this approach is that it is quite robust, since there are no iterative processes that have to converge. Even instability phenomena like wrinkling have been simulated by means of this procedure. The procedure is called the *static-explicit* approach in order to distinguish it from the better known dynamic-explicit approach.

This procedure has been particularly popular among Japanese researchers. A couple of recent papers on this subject are Nakamachi[8], and Kawka and Makinouchi[9].

The dynamic, explicit method Metal forming problems can generally be considered to be quasi-static problems, i.e. inertia forces do not have any major influence on the processes. All

the previous approaches can be considered as ‘natural’ in the sense that they are quasi-static. Despite this fact an approach, in which the problem is treated as a transient, dynamic one, has become the most popular method in later years. This particular type of method has previously been used to simulate highly transient problems like explosions, projectiles penetrating targets, automobile crashes, and so on.

The reasons for using a method like this in metal forming problems are twofold: The method is extremely robust, and it is very efficient, especially for large-scale problems.

The discretized dynamic equations are integrated by the central difference explicit time integration scheme. Furthermore, lumped mass matrices are used, which implies that the mass matrix is diagonal, and no system of equations has to be solved. The critical time step is approximately equal the time for a bending or compression wave to travel through the smallest element in the mesh. A typical time step in a sheet forming analysis is therefore of the order of a microsecond. The number of time steps in a typical sheet forming simulation is normally several tens of thousands.

Other advantages of the dynamic, explicit method are, for instance, that, because of the small time steps, the kinematic and contact conditions become very simple. The memory and data storage requirements are, furthermore, relatively small. The method is well adapted for vectorization and parallelization.

In the dynamic explicit method the computing time is directly proportional to the duration of the analyzed event. In order to speed up the computations it is customary to use a fictitious time scale and/or a fictitious density. It is, however, essential to control that the inertia forces do not influence the solution.

A majority of the most popular commercial codes for sheet metal forming simulation are based on the dynamic, explicit method. See for instance Hallquist et.al.[10], Haug et.al.[11], Mercer et.al.[12], and Aberlenc et.al.[13].

On-step methods The so-called *one-step methods* are variants of the static-implicit method, where the complete solution is performed in one single step under the assumption of linear strain paths. The history dependency of material and contacts are thus neglected. The main advantage of these methods is of course the short computing time, which is a fraction of the one for any of the previous methods.

In spite of the considerable simplifications introduced in these methods, they still have proven useful in some applications. Especially in early phases of the tool design process, even the rough predictions from a code like this can be a valuable aid. However, in later evaluations of process and die designs more accurate methods have to be used. Recent presentations of one-step methods can be found in Batoz et.al.[14] and El Mouatassim et.al.[15].

The AUTOFORM approach The approach used in the commercial code AUTOFORM is another variant of the static-implicit method. Normally, quasi-static, implicit codes make use of direct, linear solvers. The disadvantage of such solvers is that the computing time increases roughly with the second to the third power of the number of equations, which makes them less suitable for large scale problems. Iterative solvers, on the other hand, for which the computing time increases almost linearly with the size of the problem, are inappropriate for sheet metal forming problems, since the resulting system of equations is highly ill-conditioned. A

condition for an efficient utilization of an iterative solver is that the system of equations is well conditioned.

AUTOFORM uses basically membrane element, but bending can be considered as a secondary effect. The special feature of this code is that, in each time step, the motions of the nodes perpendicular to the tool surfaces are uncoupled from motions tangential to these surfaces. In each new step a form of the sheet is first sought, that satisfies the boundary conditions determined by the tools. Thereafter equilibrium is determined iteratively. Within this process the nodes have only two degrees of freedom each - two translation components in a tangent plane to the tool surface. The resulting system of equations is well conditioned, and an iterative solver can effectively be utilized.

The advantage of the present approach is that it is highly efficient and robust. The disadvantage is that, since it is based on membrane theory, some approximations are introduced in the solution, and phenomena related to bending, such as wrinkling, cannot be directly simulated.

For a more detailed description of this approach the reader is referred to, for instance, Kubli and Reissner[16].

4 THE PRACTICAL USE OF SHEET METAL FORMING SIMULATION IN A HISTORICAL PERSPECTIVE

The bases for FE analysis of large deformation problems were not established until the mid 70's, and it was not until then the first procedures for FE simulation of sheet forming were presented. Early attempts to simulate sheet metal forming processes by means of the Finite Element method were usually based on 2D, or axisymmetric models. The 'flow' and the 'rigid-plastic' approaches were more popular than the 'solid' one, mainly because it was possible to advance the solution in much bigger increments in these approaches.

In 1978 Wang and Budiansky[17] published the first complete 3D formulation for sheet forming problems, based on a membrane formulation and a 'static-explicit' approach. The practical application of sheet forming simulation was, however, during many years hampered by too unstable numerical procedures and excessive computing times, even for very small problems.

Ten years later, in 1988, Tang et.al.[18] published results from practical applications of a code, developed at Ford, to the simulation of stamping of real 3D automotive parts. This code was based on large strain shell theory and a 'static-implicit' approach. Models with up to 400 elements were analyzed, and the reported computing time was about 20 hours.

In 1989 results from a Volvo/Control Data project was presented (Honecker and Mattiasson[19]), in which the 'dynamic-explicit' approach was evaluated in application to sheet metal stamping. The results from this study were very promising. Problems with up to 10,000 shell elements could be solved within 1.5 hour on a super computer. Also the robustness of this approach was found to be widely superior to that of any other method.

From that time the practical utilization of sheet forming simulations within the industry has shown an explosive development. Most companies within the automotive industry are today

performing sheet stamping simulations on a regular basis. Dynamic explicit codes, such as LS-DYNA, PAM-STAMP, OPTRIS, ABAQUS/Explicit, and others, are dominating the software market. Exceptions can be found in Japan, where also a couple of codes based on the 'static-explicit' approach have found some industrial usage. The highly specialized code AUTOFORM (see Sect. 3.2) is also widely used, often as a complement to other codes. Various one-step codes are frequently used as preliminary design tools.

There are several reasons for the breakthrough of simulation aids in the sheet forming industry in later years. One reason is of course the development of efficient and robust simulation methods. Another equally important factor is the rapid development of computer hardware, which makes it possible for most companies to perform simulations of complex production parts in reasonable time and to a reasonable cost. However, the forming simulation is just one activity in a chain of activities. A necessary condition for the success of forming simulations has also been the rapid development of the softwares used before and after the forming simulation in this chain of actions. For instance, a necessary condition is the availability of CAD systems in which the geometry of the products can be numerically described and easily modified. Another necessary condition is the availability of efficient tools for creating FE meshes on the CAD surfaces. Finally, the development of computer graphics and efficient post-processors make it possible to easily evaluate the huge amount of output data from the simulation codes.

5 MATERIAL MODELING

5.1 Introduction

The cold rolling of the sheet material generates crystallographic textures, which is observed as a mainly orthogonal plastic anisotropy. The anisotropy normal to the sheet surface is known to be the most significant one, and is known as *normal anisotropy*. If also the anisotropy in the plane of the sheet is considered, the term *planar anisotropy* is used.

The level of the anisotropy is usually characterized by the *plastic anisotropy parameter* R, defined as the relation between plastic strain rates in the width and thickness directions, respectively, in a uniaxial tension test, i.e.

$$R = \frac{\dot{e}_b^p}{\dot{e}_t^p} \tag{5}$$

Normally, tension tests are performed on sheet strips cut from the blank in three different directions: 0^o, 45^o, and 90^o to the rolling direction. When only normal anisotropy is considered, an average value of the anisotropy parameters in three directions is used:

$$R = \frac{R_0 + 2R_{45} + R_{90}}{4} \tag{6}$$

Normally $1 < R < 2$ for steel, and $R < 1$ for aluminium. Below some of the most common yield surfaces for normal and planar anisotropy will be reviewed.

5.2 Yield criteria for normal anisotropy

In 1948 Hill[1] presented his classical quadratic yield function for three dimensional, orthogonal, anisotropic plasticity. Especially in its plane stress, normal anisotropic form, it is the, without comparison, most widely used yield criterion for sheet forming applications. The expression for the effective stress is given in Eq. (7) (compare Eq: (1))

$$\bar{s}^2 = s_1^2 + s_2^2 - \frac{2R}{R+1} s_1 s_2 = \\ = s_x^2 + s_y^2 - \frac{2R}{R+1} s_x s_y + 2\frac{2R+1}{R+1} t_{xy}^2 \tag{7}$$

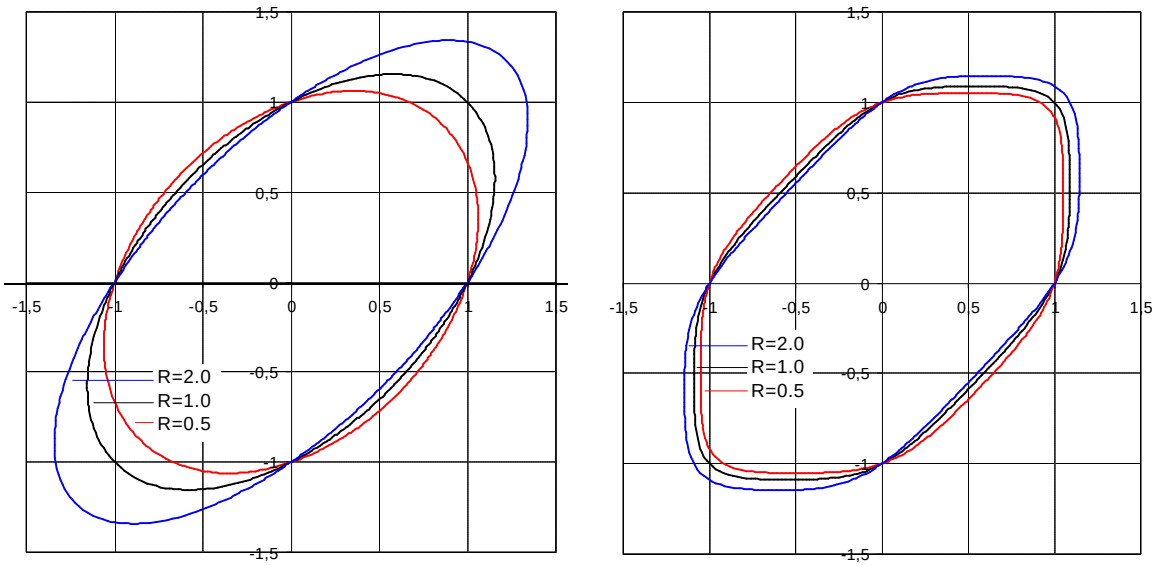


Figure 3 (a) The Hill'48 yield function, and (b) Hosford's yield function for m=8

Hill's yield surfaces for different values of the anisotropy constant R are shown in Fig. 3a. Hill's yield condition has proved to yield good result for mild steel. However, for high strength steel qualities, and especially for aluminium, it fails in providing satisfactory results.

Another yield function, suggested by Hosford[20], has been shown to yield excellent fit to crystallography-based yield surfaces for values of the exponent m in the range 6-8 (see Eq. (8)). Hosford's yield surface for different values of R is displayed in Fig. 3b. The expression for the effective stress is shown in Eq. (8). Hosford's criterion, can be shown to reduce to Hill's quadratic yield function in the special case when R=2.

$$\bar{s}^m = a_1\left(|s_1|^m + |s_2|^m\right) + a_2|s_1 - s_2|^m \tag{8}$$

$$a_1 = \frac{1}{1+R}; \qquad a_2 = \frac{R}{1+R};$$

It has been observed in numerous experiments that the shapes of the yield surfaces for metallic materials can be found somewhere between two extremes, represented by the yield surfaces of Tresca and von Mises, respectively. It is interesting to note that Hosford's yield surface, for increasing value of the exponent m, approaches Tresca's yield surface.

5.3 Yield criteria for planar anisotropy

A great number of yield criteria for planar anisotropy have been presented. Here a couple of the most well known criteria will be presented.

Barlat and Lian[21] have proposed a yield criterion, which can be viewed as a generalization of Hosford's criterion for normal anisotropy to the more general planar anisotropic case. An advantage of this criterion is that the anisotropy properties can be represented by parameters obtained in simple standard tests. The effective stress in the Barlat-Lian criterion can be expressed as

$$2\bar{\sigma}^m = a\left|K_1 + K_2\right|^m + a\left|K_1 - K_2\right|^m + c\left|2K_2\right|^m \tag{9}$$

$$K_1 = \frac{\sigma_x + h\sigma_y}{2};$$

$$K_2 = \sqrt{\left(\frac{\sigma_x - h\sigma_y}{2}\right)^2 + p^2\tau_{xy}^2};$$

Note that this expression corresponds to Hosford's yield criterion with the normal stress in the y-direction weighted by a factor h, and the shear stress weighted by a factor p. The material constants a, c, and h can be expressed in terms of R-values in the 0^o- and 90^0-directions. The parameter p cannot be calculated directly, but has to be solved for iterativly from an equation involving the anisotropy parameter R in the 45^o direction. An extension of Eq. (9) to three dimensional, orthotropic plasticity has been proposed by Barlat et.al.[22].

Karafillis and Boyce[23] used the "mapped stress tensor" concept to derive a three dimensional, anisotropic material model. In their model they introduced a linear transformation tensor acting on the stresses σ_{ij} in the real material. The transformation tensor "weights" the different stress components of the anisotropic material. The weighted stresses can be considered to act on a corresponding, fictitious, isotropic material. For the case of an isotropic material, the transformed stress tensor will be equal to the deviatoric stress tensor acting on the real material. The transformed stress tensor is called the "isotropy plasticity equivalent (IPE) deviatoric stress tensor". The transformation can be written

$$\tilde{S}_{ij} = L_{ijkl}\,\sigma_{kl} \tag{10}$$

The isotropic yield function, corresponding to the stress state $\tilde{S}_{ij}$, is prescribed and the elements of the transformation tensor, describing the anisotropy of the material, are

determined from a suitable set of experiments. In the present case the isotropic yield function is general enough to able to describe both the lower (Tresca) and the upper bounds, existing for isotropic yield functions.

6 PREDICTION OF FORMING DEFECTS

6.1 Introduction

Current methods and codes for sheet metal forming are quite successful in predicting parameters, which are related to the deformation of the sheet material, for instance strain distributions, thinning, and draw-in of the blank edge. The forces acting in the interfaces between the blank and the tools can usually also be predicted with satisfactory precision. However, some forming defects like rupture, springback, and surface deflections, can unfortunately not always be predicted with the desired level of accuracy.

Much of the modern research on sheet metal forming simulation is consequently devoted to these particular issues. It is, however, the object of the current presentation to describe the methods that are in practical use today to predict these defects.

6.2 Prediction of rupture

The risk for rupture in the material is usually evaluated by means of a so-called Forming Limit Diagram (FLD). This is an experimentally determined curve in the principal strain plane, showing combinations of principal strains leading to rupture. In these experiments rectangular sheets with different widths are stretched over a hemispherical punch until rupture occurs. Every single width of the sheet specimen corresponds to a unique linear strain path up to failure, and gives one point on the FLD.

The risk for failure is normally evaluated in the post-processing of the results from the simulation code, but the FLD can also be built in the material model as failure criterion. Critical zones in the formed part can be detected by visualising a “failure index”, defined as

$$c = \varepsilon_1 / \mathrm{fl}(\varepsilon_2) \tag{11}$$

where $\mathrm{fl}(\varepsilon_2)$ is the forming limit curve viewed as a function of the minor principal strain. This index indicates rupture when $c \geq 1$. In Fig. 4 this index is visualised as colour fringes for a formed panel. As can be seen a critical area is detected and is marked by red colour.

If the strains in the middle surface of every element in the critical area are plotted in a principal strain diagram this results in a diagram like the one in Fig. 5. Some strain points in this diagram are situated above the forming limit curve, indicating material failure in the corresponding elements.

A lot of criticism can be raised against the use of FLDs as failure criteria. There is first of all a big uncertainty about the exact appearance of the forming limit curve itself, since this is highly dependent on the test procedure. Secondly, the FLD is created from linear strain paths, while the strain paths leading to failure in the actual forming operation very seldom are linear. It has in fact been shown that the limit strains are highly dependent on the strain path. This

deficiency of the conventional failure evaluation procedure is especially evident for multistage forming processes.

Current research on methods for rupture prediction is therefore concentrated on finding procedures that can handle nonlinear or broken strain paths. Stress based forming limit concepts and damage mechanics models are example of attempts in that direction.

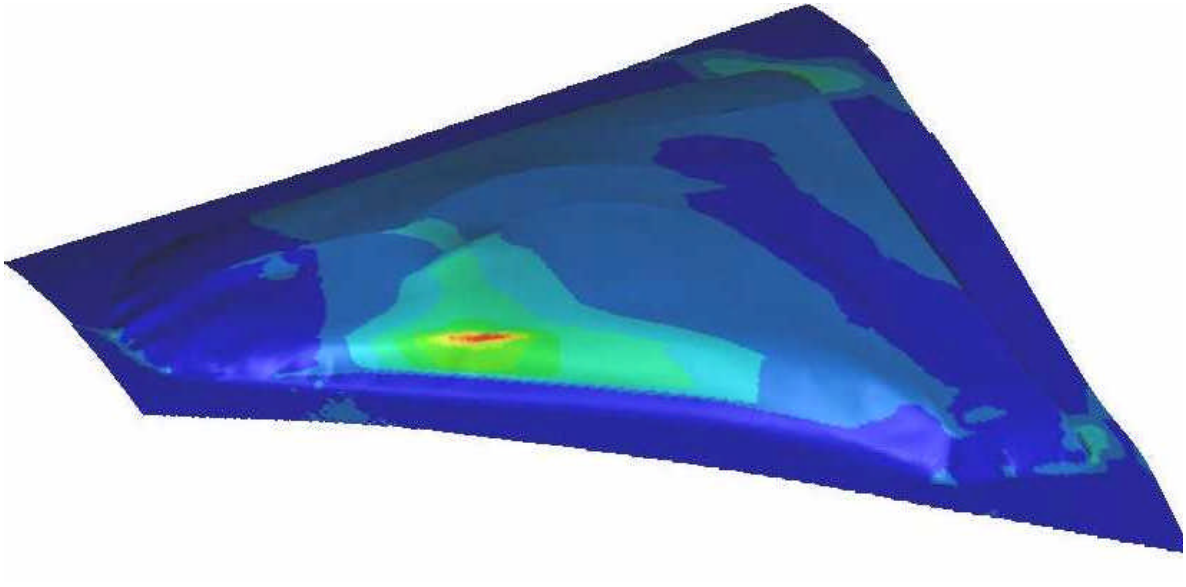


Figure 4 Colour fringes indicating ”failure index”. Red colour indicates rupture.

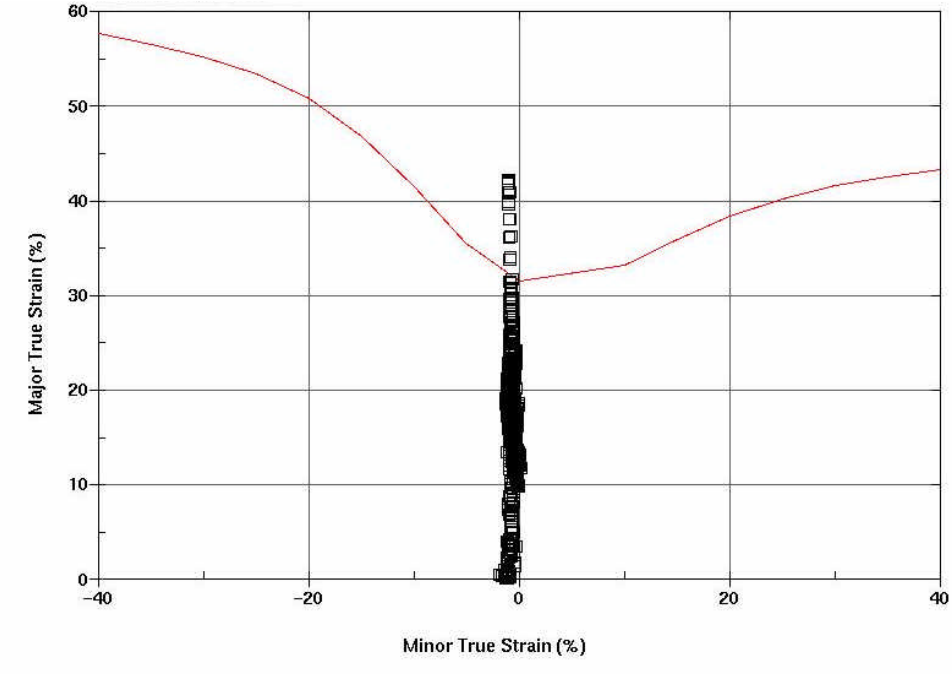


Figure 5 Forming limit diagram for the critical zone visualised in Fig. 4.

6.3 Springback

Springback analyses by means of dynamic, explicit codes must, in contrast to forming simulations, be performed in a real time scale. The normal procedure is then to apply boundary conditions so that rigid body motions are prevented, remove the tools instantaneously, apply a suitable amount of damping, and then let the work piece vibrate freely until a static equilibrium is reached. The drawback of such a procedure is that, first of all, it is very difficult to estimate what amount of damping should be applied without first doing a separate eigenfrequency analysis, and, secondly, that the time for reaching a static equilibrium many times can be several times longer than the time needed for the actual forming operation. For this reason the static-implicit method is preferred for springback analyses.

A condition for an accurate springback analysis is that the calculated stresses after the completed forming simulation are correct. It has, however, been shown that it is much more difficult to obtain accurate stresses than accurate strains.

In connection to the NUMISHEET'93 conference a benchmark test was set up, which aimed at letting the participants determine the springback in a deep-drawn, U-shaped sheet strip, both experimentally and/or numerically. The numerical benchmark results were, however, very disappointing, showing a great scatter among the different participants. Most codes seemed to strongly underestimate the springback.

Later on the author and colleagues, Mattiasson et.al.[24], did reanalyse the present problem in order to find out the causes of the inaccurate springback predictions. Especially the influences of various model parameters on the resulting stress state after completed forming were studied. In Fig. 6 the longitudinal stress history during the forming operation in a point on the outer surface of the sheet strip is displayed. The influence of the mesh size in the sheet is studied, and results are shown for element sizes 3.0 mm and 0.5 mm. For the larger element size a pronounced relaxation of the stresses, after the point in question has left the draw radius, can be observed. The results for the finer mesh, which represent a converged solution with respect to the element size, do not show this stress relaxation. In Fig. 7 the geometry of work piece after springback is displayed for various element sizes.

The observed stress relaxation phenomenon could be explained by the small variations in strains in the vertical part of the work piece, which are caused by the basically flat elements in the sheet slipping over the draw radius.

The referred study showed, thus, that the main reason for the inaccurate springback results was the use of a too coarse mesh in the sheet. In fact, an extremely fine mesh was needed in order to get a converged solution. There were, however, a number of other factors that had substantial influence on the results. For instance, it was shown to be very important to include the Bauschinger effect in the modelling of the material hardening. Furthermore, the fictitious process time used in a dynamic-explicit method should be at least twice the time normally used, when an accurate solution for strains is of primary interest.

Even though the above observations have been considered, the agreement between measured and calculated springback for complex parts have in many cases been poor. This

indicates the problem of springback is not yet fully understood, and that this should be a focused are for current research.

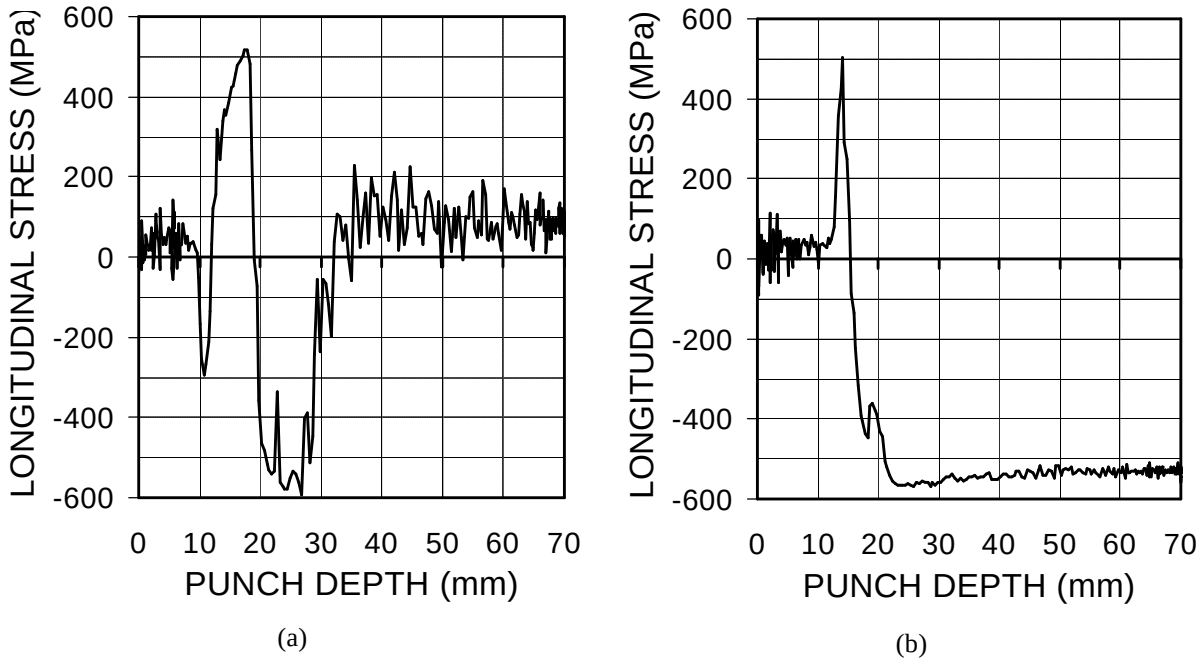


Figure 6 Stress history at a point on the outer surface of a U-shaped deep drawn sheet strip (from Mattiasson et.al.[24]). (a) Element size 3.0 mm, (b) Element size 0.5 mm

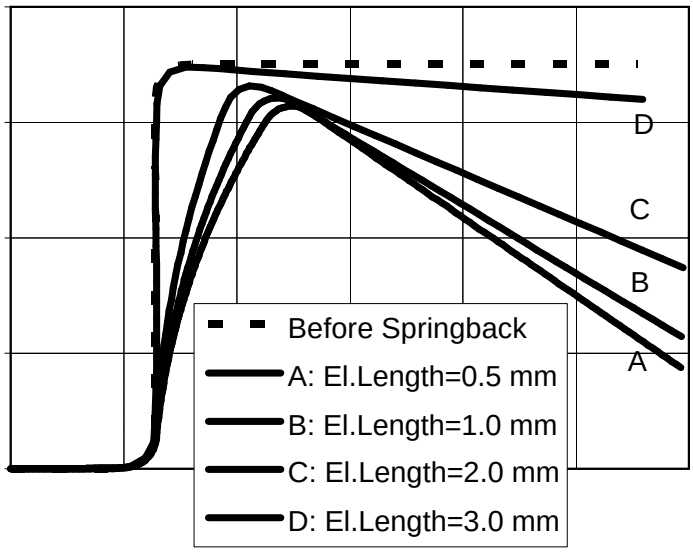


Figure 7 Geometry after springback for different element sizes (from Mattiasson et.al.[24])

7 SUMMARY

Sheet metal forming simulation is today used by routine by most car manufacturers and major tool makers. Simulations of varying complexity are performed in different phases of the forming process development. One-step codes are mainly used in the early product design stage to evaluate manufacturing feasibility. The advantage of one-step codes is the short turn around time. Computing time as well as the time needed for data preparation are considerably shorter than for incremental codes.

More thorough analyses are performed by means of incremental codes to support the die and process design. Today the software market for this type of codes is dominated by codes based on the dynamic-explicit method. The computing time for complex production parts is typically several hours.

There are some areas of sheet forming simulation for which there exist particular needs for further research and development. This concerns especially detection and evaluation of certain types of forming defects, such as rupture, springback and surface deflections.

REFERENCES

[1] R. Hill, "A theory of the yielding flow of anisotropic metals", *Proc. R. Soc. London*, **193A**, p281 (1948)

[2] E. Onate and C. Agelet de Saracibar, "Alternatives for finite element analysis of sheet metal forming problems", *NUMIFORM'92* (Eds.: Chenot, Wood and Zienkiewicz), Sophia Antipolis, A.A. Balkema (1992)

[3] N.-M. Wang, "A rigid-plastic rate sensitive finite element procedure for modeling sheet forming processes", In *Numerical Analysis of Forming Processes* (Eds.: Pitman, Zienkiewicz, Wood, and Alexander), John Wiley & Sons, p 117 (1984)

[4] C.H. Toh and S. Kobayashi, "Deformation analysis and blank design in square cup drawing", *Int. J. Mach. Tool Des. Res.*, **25,** p 15 (1985)

[5] Y.Geramain, K. Chung, and R.H. Wagoner, "A rigid-viscoplastic Finite Element program for sheet metal forming analysis", *Int. J. Mech. Sci.*, **31**, p 1 (1989)

[6] R. Sünkel, C. Pautsch, K. Roll, R. Toderke, F. Fuchs, and V. Steininger, "Numerical simulation of metal forming processes in industry using INDEED", *NUMISHEET'96* (Eds.: J.K. Lee and R.H. Wagoner), Dearborn, Michigan (1996)

[7] S.C. Tang and Y. Hu, "Quasi-static analysis of sheet metal forming processes on a parallel computer", *NUMIFORM'95* (Eds.: Shen and Dawson), Ithaca, N.Y., A.A. Balkema (1995)

[8] E. Nakamachi, "Sheet forming process characterization by static-explicit anisotropic elastic-plastic finite element simulation", *NUMISHEET'93* (Eds.: Makinouchi, Nakamachi, Onate and Wagoner), Isehara, Japan (1993)

[9] M. Kawka and A. Makinouchi, "Finite element simulation of sheet metal forming processes by simultaneous use of membrane, shell and solid elements", *NUMIFORM'92*

(Eds.: Chenot, Wood and Zienkiewicz), Sophia Antipolis, A.A. Balkema (1992)

[10] J.O. Hallquist, B. Wainscott, and K. Schweizerhof, "Improved simulation of thin sheet metal forming using LS-DYNA3D on parallel computers", *NUMISHEET'93* (Eds.: Makinouchi, Nakamachi, Onate and Wagoner), Isehara, Japan (1993)

[11] E. Haug et.al., "Numerical simulation of industrial sheet forming processes with PAM-STAMP", *4th Cars/Trucks Symposium*, Schliersee (1995)

[12] C.D. Mercer, J.D. Nagtegaal, and N. Rebelo, "Effective application of different solvers to forming simulations", *NUMIFORM'95* (Eds.: Shen and Dawson), Ithaca, N.Y., A.A. Balkema (1995)

[13] F. Aberlenc, J.-L. Babeau, and P. Jamet, "OPTRIS: The complete simulation of the sheet metal forming", Ibid.

[14] J.L. Batoz, Y.Q. Guo, and F. Mercier, "The inverse approach including bending effects for the analysis and design of sheet metal forming parts", Ibid.

[15] M. El Mouatassim, B. Thomas, J.-P. Jameux, and E. Di Pasquale, "An industrial finite element code for one-step simulation of sheet metal forming", Ibid.

[16] W Kubli and J. Reissner, "Optimization of sheet metal forming processes using the special-purpose program AUTOFORM", *NUMISHEET'93* (Eds.: Makinouchi, Nakamachi, Onate and Wagoner), Isehara, Japan (1993)

[17] N.-M. Wang and B. Budiansky, "Analysis of sheet metal stamping by finite element method", *J. Appl. Mech. Trans. ASME,* **45**, p73 (1978)

[18] S.C. Tang, R. Ilankamban, and P. Ling, "A finite element modeling of the stretch-draw forming process", *SAE-paper* 880527 (1988)

[19] A. Honecker and K. Mattiasson, "Finite element procedures for 3D sheet forming simulation", *NUMIFORM'89* (Eds.: Thompson, Wood, Zienkiewicz and Samuelsson), Fort Collins, A.A. Balkema (1989)

[20] W.F. Hosford, "Comments on anisotropic yield criteria", *Int. J. Mech. Sci.*, **27**, p423, 1985.

[21] F. Barlat and J. Lian, "Plastic behavior and stretchability of sheet metals. Part I: A yield function for orthotropic sheets under plane stress conditions", *Int. J. Plasticity*, Vol 5, pp 51-66 (1989)

[22] F. Barlat, D.J. Lege, and J.C. Brem, "A six-component yield function for anisotropic materials", *Int. J. Plasticity*, Vol 7, pp 693-712 (1991)

[23] A.P. Karafillis and M.C. Boyce, "A general anisotropic yield criterion using bounds and a transformation weighting tensor", J. *Mech. Phys. Solids*, Vol 41, 12, pp 1859-1886 (1993)

[24] K Mattiasson, A Strange, P Thilderkvist and A Samuelsson, "Simulation of springback in sheet metal forming", *NUMIFORM'95* (Eds.: Shen and Dawson), Ithaca, N.Y., A.A. Balkema (1995)

Digital Manufacturing in Press Part Production

Dr. Schiller, Prof. Dr. Roll, Dr. Wöhlke, Mr. Wiegand

DaimlerChrysler AG, Production Planning Mercedes-Benz Passenger Cars, Germany

1 Challenges

In view of the challenges faced by automobile manufacturers today, there are a number of factors that need to be considered. At least in the triad markets (Europe, the USA and Japan), manufacturers face increasingly global competition on saturated markets. This development is also expressed in steadily growing pressure for consolidation and concentration in the automobile industry. Concentration of course not only affects vehicle producers but also component manufacturers. In this context, component manufacturers are focussing more and more on complete packages in the value addition chain. As a result, new key players have emerged; to a growing extent, they are also increasing the competition in the automobile industry.

1.1 Product offensive - effects

In spite of these challenges, DaimlerChrysler shows that the company has already succeeded in adapting to new requirements and in continuously boosting production figures in the past. Over the past five years, sales have almost doubled. One of the key factors in this development has been the product offensive which was successfully launched at the beginning of the 1990's. A large number of attractive model series and product variants have been developed in the course of this offensive (see Fig. 1).

Fig. 1: A large number of attractive models have been developed as a result of DaimlerChrysler's "product offensive" during the 1990s.

However, this gratifying development has had dramatic effects on production plants and on the engineering and design departments. Whereas in the past it was only necessary to design a new vehicle at intervals of 2 to 3 years, this interval has now been reduced to 3 to 4 months. It has almost become normal not only for one model series but also for several vehicle projects to be in the design and start-up process at the same time. In addition, pressures on costs are steadily growing.

1.2 The MB Development System (MDS) process model

However, it is only possible to tap the potential for reducing design and start-up times if the underlying processes are precisely defined and a binding definition of the links between these processes is available. In this context, DaimlerChrysler already started work on the Mercedes-Benz Development System (MDS) six years ago. This process model describes the content of the individual phases, from the strategy, technology and vehicle phases through to series production, and precisely defines the time links between them. The objective is to record as comprehensively as possible all the major activities involved in process phases and to set out the links between the individual units concerned, e.g. between development and the various engineering departments. Compliance with the required degree of process and product maturity is monitored at various milestones, referred to in the model as quality gates.

Conventional development and engineering methods alone are inadequate for shortening development processes and design times to the extent required. In the future, the production engineering departments will also need to stronger implement digital planning and review methods similar to the Digital Mock-up (DMU) review of 3D product models in the development departments and to apply these methods on a consistent basis. This applies especially to processes which are on the critical path within overall project planning or would result in considerable cost and effort in the event of any changes. A typical example are processes for the design and review of press tools for the production of body parts which depend considerable on the manufacturing time of the tool builder. These processes have a considerable impact on the start of production (SOP) and the start-up curve of a new vehicle project. In order to shorten development times for press tools more drastically in the future, DaimlerChrysler is introducing an engineering process for sheet metal part production with digital support.

With respect to the MDS process outlined above, the following activities offer considerable potential for the use of digital design methods in tool production for body and structural components (see Fig 2):

- If a **manufacturing feasibility review** is carried out at an early stage in the technology phase, the feasibility of manufacturing the part in terms of forming operations can be ensured on the basis of an initial simulation without any need for the production of costly prototype tools or complex testing.
- Before the tool design specification has been issued, forming geometries for the various process stages and the tool designs derived from these geometries can be changed on the basis of simulations until the ideal part is produced. This approach results in a further reduction in **engineering time** combined with a significant **increase in maturity**.
- As part of an **overall review** (carried out prior to commissioning), the fit and function of the press tool design can be verified by a virtual inspection process while

design is in progress. The commissioning can then be started when this review has been completed.

- The digital models can also be used for **start-up support** (prior to ramp-up), for the advance training of operation and maintenance personnel and for the offline generation of the control programs actually to be used for production. This results in a significant shortening of the start-up phase following the introduction of new tools.

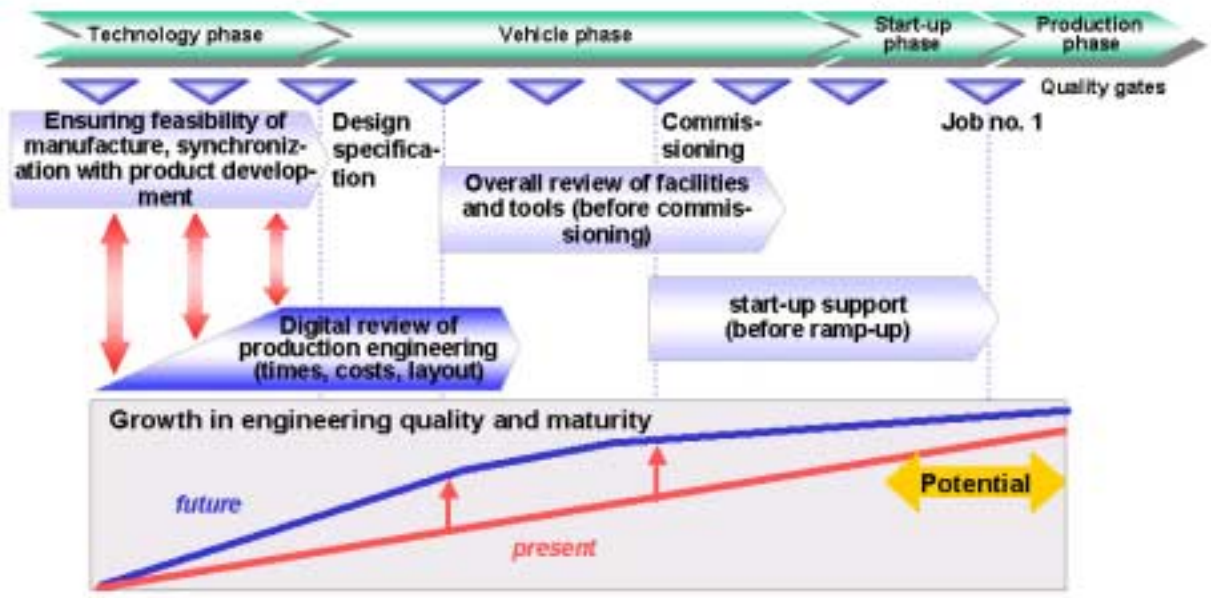


Fig. 2: MDS Process Model – Potential of Digital Planning

The workflow including the various process stages from **methods planning**, via **forming simulation** and **tool design** to **press line/shop simulation** and the use of the digital tools and systems concerned are described in greater detail below.

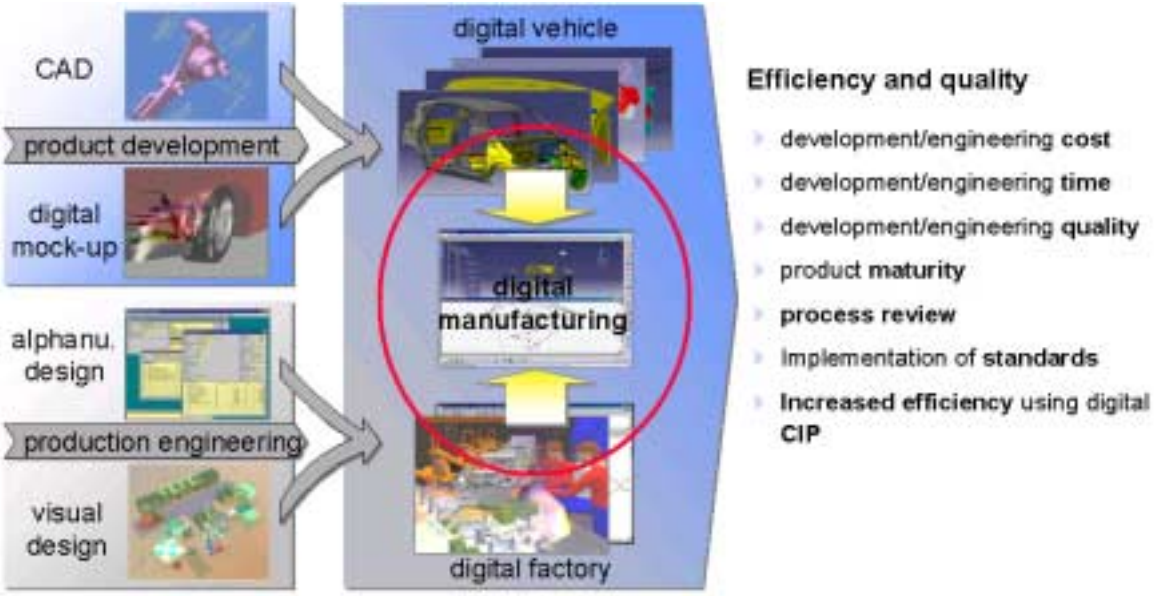


Fig. 3: Interaction between Product Development and Production Engineering

As a result of the consequently shortening of product development and production engineering periods, production engineering and development have increasingly be-

come parallel processes (see Fig. 3). From the production planning point of view, production requirements must therefore be taken into consideration as early as possible in the development process, at a point where product data are only available in digital form, as CAD models.

2 The vision of the digital factory

In this context, the DaimlerChrysler vision of the digital factory is as follows: in the future, no production facility will be designed, constructed or commissioned without a full review carried out using digital design methods. The review will cover the entire factory and all buildings as well as individual units such as shops, production lines, cells and manual work stations. Individual tools, operational steps and technical operations such as welding, bolting or adhesive bonding will also be included. It will be important not to neglect the human factor, both in combination with tools and machines and with respect to ergonomic aspects such as physical loads.

If this vision is viewed in isolation with respect to individual levels, it may not seem to be very innovative as various digital design and simulation methods have been used for individual process stages in the past. However, it represents a considerable challenge if the entire system including all the individual aspects is to be linked in the form of continuous workflows with access to a central data management system. For the pressed part workflow described below, this applies in particular to the process engineers, engineering departments and software partners involved in the process.

2.1 Process engineering activities

The objective of the engineering process is to design and realize the production facility required on the basis of the available product data. This process is currently supported by a wide range of software systems. Normally, there is only a file-based data exchange mechansim between the various systems used. The data are stored in a number of different local databases and often need to be transferred from one system to the next manually. As a result, databases are often not up-to-date and complex and time-consuming data compilation and reconciliation processes are unavoidable. Keeping the present situation in mind, the following sections consider the directions in which the methods of the digital factory have to be developed in order to succeed in the future, with special reference to planning of press part production.

2.2 Main approaches for the development of the digital factory

In the opinion of DaimlerChrysler, the new digital design methods of the digital factory will need to be based on the following four main approaches (see Fig 4). It will be necessary to apply these approaches consistently if the potential time and cost savings referred to above are to be realized in the process as a whole.

1. Initially, **standards** and production principles in accordance with the Mercedes-Benz Production System (MPS) will need to be defined and systematically supported.
2. In the future, **data integration** will be necessary in order to replace the wide variety of individual databases currently used by a few data management systems. Systems must be designed in such a way that each data record only needs to be recorded and saved once and the supplier of the data remains responsible for updating it. This will apply to all data, including 3D product and factory data, tool and equipment data records, process and production plans and simulation results.

Of necessity, this will call for certain changes in the approaches and working methods of development and engineering staff. In the future, it will be necessary to save incomplete data and interim versions on the system and not just complete, reviewed results. In addition, all the information saved will need to be available throughout the world.

3. Processes will need to be defined and integrated in the form of **workflows** so that the sequential working methods currently used can be replaced by a form of meshed cooperation including revision management.
4. The **automation** of repetitive routine design tasks will relieve the workload on production engineers and ensure further benefits.

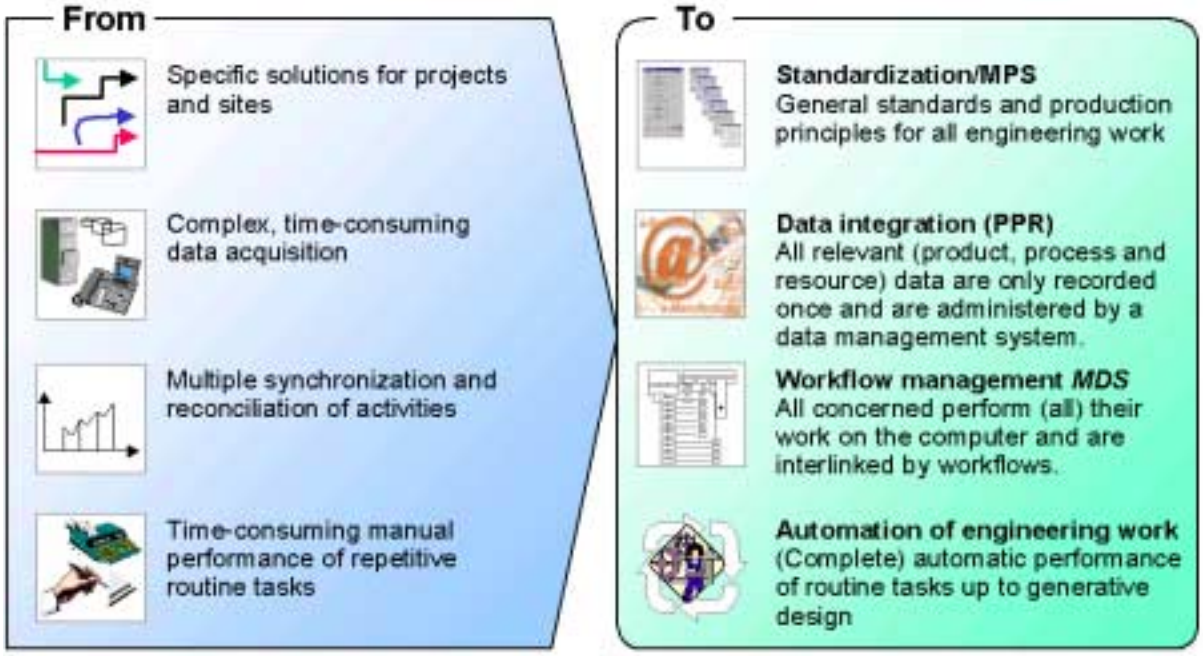


Fig. 4: Main Approaches of the Digital Factory

3 Sheet metal part production workflow

These aspects are illustrated below on the basis of the workflow for sheet metal part production. The objective is also to demonstrate how an ideal development and review process for press tool design can already be supported by digital methods today. An important approach is the use of deep drawing simulation early in the technology phase in order to verify forming feasibility and define forming geometries. In this way, the production of a component in the press line can already be simulated over several stages from the design model through to the actual design of the press tools before the first prototype is produced (see Fig. 5).

- Starting with the **product data** (CAD geometry and product structure) from vehicle development, the production-specific attributes such as material, sheet thickness, etc., required as a basis for subsequent process stages are added.
- In **methods planning,** the sequence of operations needed for the forming of the component is defined and the necessary **additional design work** is performed; in other words, active areas are added to the CAD model of the finished product. These additions include the geometry of the holders required to fix the part in position during pressing and the resulting punch and die geometry.

- After forming geometries have been defined and the FEM net has been automatically generated, the flowing behavior of the sheet metal during the deep drawing process is determined by a **forming simulation.** The results of the simulation are visualized for assessment. This approach indicates points where the sheet may become too thin, leading to cracks, and areas of excessive thickness where buckling could occur. In addition to the deep drawing stage, all the **subsequent stages**, such as cutting, hemming, folding, etc., as well as springback effects on the component can now be simulated. In this way, it is possible to optimize the entire production process in advance.
- The next stage following the completion of the manufacturing feasibility review is **tool design**; in this stage, 3D solid models of the press tools are generated on the basis of the forming geometry defined in previous stages. These models can be reviewed by virtual inspection using a **press line simulation**, in other words a DMU study of the tool. This allows the mechanical elements required, such as vacuum holders, supports, etc., to be defined and the control programs for the press to be generated. At the same time, it is possible to define and optimize the physical properties of the tool using strength calculations (FEM net generation and simulation).
- **Press shop simulation** is an overall review process which allows the investigation and verification of deployment planning and various alternative logistics configurations (for blank supply and finished part handling). Using the results, throughputs and pressing rates can be optimized at an early stage. It is even possible to test future maintenance processes on the digital model.

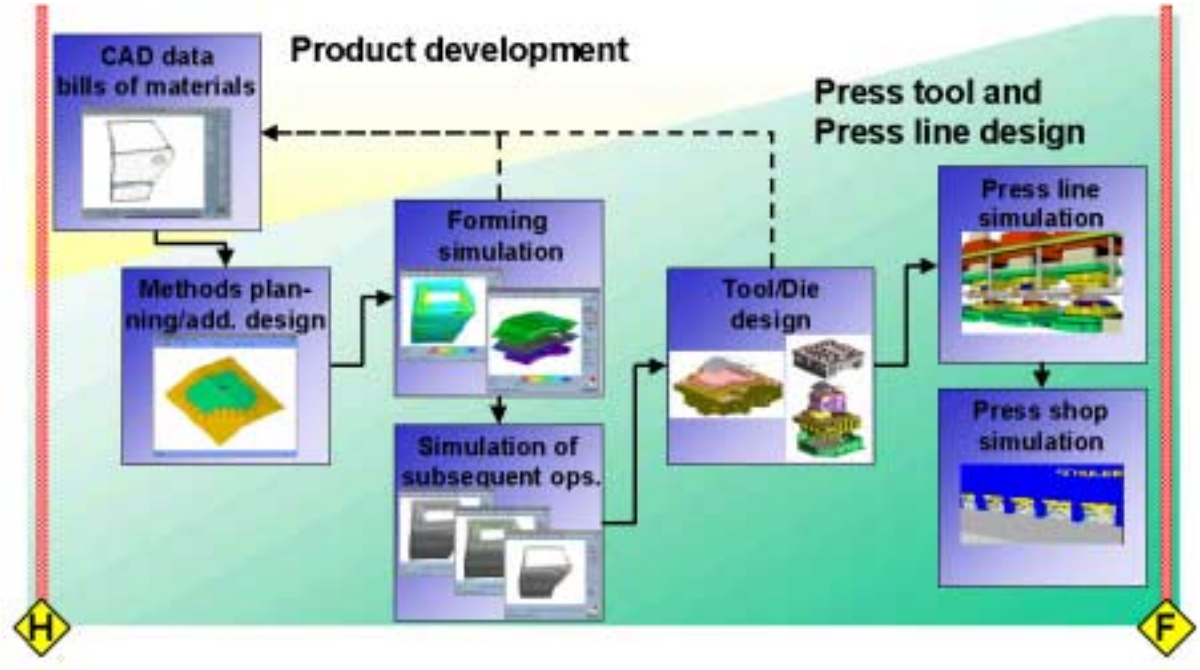


Fig. 5: Sheet Metal Part Production Workflow

Commercial software solutions are now available to support all the stages mentioned above, allowing a continuous digital workflow. You could say that a part runs through the digital factory before it is actually physically produced. It is optimized and improved until it is ready for production by the physical factory. This approach has considerable benefits:

- Changes in basic planning data and design improvements can be implemented cost-effectively and reviewed digitally. Especially, this process can not only be

carried out stepwise and in one direction. By simply varying the parameters used, a large number of scenarios can be generated and compared.

- In the workflow system, requests for changes can be transmitted to upstream and downstream process stages using a defined change management system with suitable communications mechanisms. Defined standard elements, the data integration and documentation procedures described above and digital design methods are all used in a tightly meshed network for the individual process stages.
- The most important benefit is that digital methods allow engineering and planning know-how to be integrated into the development process at an earlier stage.

3.1 Early DMU studies

Using CAD systems and simulation combined with large-screen projection systems, allows at an early stage of the project the presentation of the actual design status of forming tools in 3D with a scale of 1:1. This approach, which depends on the capability of 3D modeling and VR visualization of the press tools, allows a continuous verification of engineering and planning progress and significantly shortens development times. Apart from virtual tool inspection, also packaging studies are possible. These studies, based on 3D models, determine whether there is sufficient space for the installation of all the parts and assemblies required. In tool production and press line design, these problems arise in connection with the design and review of press tools for sheet metal part production. Production engineers are not only interested in verifying the feasibility of producing all the parts and tools required and the compatibility of the parts and tools but also in the following questions:

- whether the components can in fact be installed in the space available,
- whether certain tools and equipment items are accessible,
- what is the ideal installation and production sequence.

3.2 Potential of digital methods

Previously, it was only possible to answer such questions at a much later stage by producing prototypes or even by trial and error on the physical press. At that stage, parts had already been produced and tools ordered; any changes required were both costly and time-consuming. With reference to current cost levels, the use of digital methods in press line design and tool production offers considerable **savings potential**.

Especially **tool modification costs** incurred as a result of the redesign of a part or the production of a completely new part can be significantly reduced by virtual tool inspection on a 3D model. As a result, the ***start-up costs*** of the tool in the press shop can also be reduced, as it is possible to raise the overall quality and maturity of the tool to series production levels at a very early stage. Part of the prototype tests oin the ty-out presses can be omitted by using digital planning methods, facilitating production engineering and capacity deployment in the pressing shop. Some savings will also be possible in terms of **capital expenditure;** there are already indications that the hardware simulators now needed in press shops will no longer be required in the future if press line simulation is used at an early stage in the design process. The potential benefits of digital engineering and simulation systems in press shop design are already evident, even without considering the shortening of the process.

4 Examples of the four main approaches

With reference to the four main **approaches** for digital methods, we would now like to discuss the individual aspects on the basis of the development and design of pressing tools and facilities for sheet metal part production.

4.1 Standardization

One example of standardization in connection with digital press shop planning are **preconfigured digital models** of press lines which are defined independently from the press tools by the press line manufacturer. In combination with predefined standard modules for the mechanical equipment, these preconfigured packages allow a very rapidly model generation process of fully-equipped press lines. These modules can then be used for the verification and optimization of the press tools, which are also designed by using predefined components. Design standards, reinforcement structures etc. can also be stored in the form of rules, allowing the generation of easily adaptable ***intelligent design modules*** for the engineering departments. Such standard elements can be easily and rapidly combined to develop the press tools and mechanical equipment required for producing a specific sheet metal part at a very early stage in the project. In addition, the cost benefits of standardization effects are substantial.

4.2 Data integration

In this context, data integration does not mean that all the data are stored in one large database but that the three engineering points of view which are relevant to the data: the **P**roduct, **P**rocess and **R**esource-oriented views - or PPR – are administered jointly by a smart data management system. This includes all the data generated from development through production engineering to production (in the pressing shop). It is clear that data from production must be transferred to the engineering and development departments, for example in order to take changes or optimizations into account.

In the case considered here, data from a variety of individual sources are needed for tool simulation and design. These include:

- tables of material data and parameters
- CAD models (surfaces) containing component data and active areas
- forming geometries and derived FEM models (networks)
- simulation results (ASCII or binary data)
- CAD models (solids) of pressing tools and jigs
- geometric or functional models of pressing facilities and machines

Data integration is based on the use of CATIA data to allow continuous work on the CAD models of components, tools and machines in a closely connected way over the entire process chain. The additional design surfaces needed for verifying manufacturing feasibility and the FEM models for forming simulation are derived from these data. Together with the forming geometries and solid tool data generated, these are used as inputs for a press line simulation. All the data records referred to must be managed consistently and must be linked with each other to allow process, product and resource-oriented views.

4.3 Workflow management – use of digital methods

As digital engineering methods become more revalent, it will become necessary to think more and more in terms of workflows and to network individual tasks previously performed in sequence more closely to shorten the entire process. This approach will go beyond the digitization and optimization of individual processes and will ensure better process integration, bringing fundamental changes in working methods. The engineer will no longer be forced to search for and acquire the data required but will be informed automatically of any changes in components or tools in the process of design or development.

Both, the presented workflow approach and the networking between individual tasks will mean that changes will not only be documented but will also be made available almost on a real time basis. With respect to tool engineering, it will therefore be possible to take tool changes into account and distribute the information required in a targeted way even before the first prototype has been built. The result will be a significant improvement in the engineering maturity of the project. The quality of integrated engineering will remain at a consistently higher level, ensuring that the results required can be obtained in a significantly shorter time.

4.4 Automation of routine tasks

As in production, manufacturing engineering includes a large number of repetitively routine tasks which could be automated with system help. Together with data acquisition work, these routine tasks take up a large proportion of the engineering capacity available, leaving relatively little time for creative work. We would like to illustrate this problem using the example of press tool design.

Using classical design methods (CATIA), it currently takes between 10 and 14 days to generate additional surfaces for the blank holder on the base of the original tool data. These surfaces are required for creating tool geometries for methods planning and forming simulation. If digital factory methods are adopted and mathematical support curves are defined for the missing surfaces, the generation process itself can be performed by the computer without further intervention. This approach reduces the time needed to between one and two hours. When the forming geometries are available, the methods engineer can then concentrate on finding the best tool shape, which can then be verified by forming simulation.

A further example is automated collision testing of 3D press tools with a press line model; this supports virtual tool inspection and functional optimization. Although virtual reality methods have been introduced for visual inspection by the engineer, this process has not yet been significantly accelerated. The full benefits of digital design methods can only be tapped a planning task can be performed automatically by the computer system. With digital design, calculations can be performed virtually overnight and the engineer is then given a result list, e.g. of the press tools which could lead to collisions in the press line. This is the point where creative engineering is called to identify feasible solutions, design modifications or tool optimization.

4.5 Digital CIP

Digital design methods also allow a continuous improvement process, digital CIP to be started on the basis of digital models before the system is built and commissioned. In the case of sheet metal part production, the first optimizations for increased operating cycles can already be performed on the 3D data records and press models that follow the tool design process. Control programs for the press line can

also be generated on this basis. Together, these developments allow operations staff to be involved at an early stage and to be trained using maintenance simulations.

Currently, there is still an interim stage in this process. The digitally reviewed results of design work are implemented as realistically as possible in the pre-production series shop and the press tools are still tested and optimized on try-out presses as a sort of hardware validation prior to the start of production. This stage can be referred to as the "physical mock-up". The advantage of digital CIP is that this process can be initiated on the basis of the digital mock-up well before operation starts.

5 Qualification requirements for the digital factory

In this context, it is necessary to develop a detailed job description and qualification profile for the digital production planner of the future. This profile must then be compared with current qualification profiles in order to define appropriate training requirements. Training will not concentrate so much on the use of software tools as on the new processes, procedures and methods involved. It will be necessary to initiate a changed consciousness or a new paradigm. For example, data, including incomplete intermediate data, will need to be disclosed and accessible at a considerably earlier stage in the future. Proactive information on any changes will be required. Training on the actual systems used should only start when the trainees have understood and adopted this new approach.

Summary and outlook

We described a new engineering process for press part production which will be implemented at DaimlerChrysler. To reach the challenging goals of reducing planning time, handling of complex part geometries, reducing manufacturing costs and improving the press tool quality, the use of digital planning methods is required. Therefore, four main approaches are presented and discussed which are considered as the basics when applying techniques of digital manufacturing to a specific area of application. We illustrated the effects at the workflow of press part production where feasibility checks and review steps are ensured with the help of digital planning and simulation tools. Resulting is a steady increase in the maturity of tool design from the initial idea through to commissioning. Also, the duration of production engineering of press tools and the amount of manual re-work can be significantly reduced by using digital planning methods. DaimlerChrysler is working within the Digital Manufacturing project on the goal to implement the necessary methods and tools to support new planning workflows. So, the digital factory plays a key role in facing up the described challenges and will force changes in working practices of planning engineers.

Numerical Analysis of Quasistatic Contact Problems in Viscoplasticity

Juan M. Viaño$(^*)$
Departamento de Matemática Aplicada
University of Santiago de Compostela
15706 Santiago de Compostela, Spain
maviano@usc.es

$(^*)$ Jointly with **José R. Fernández** (Departamento de Matemática Aplicada-University de Santiago de Compostela-Spain) and **Mircea Sofonea** (Laboratoire de Théorie des Systèmes-University of Perpignan-France

Keywords: Contact, error estimate, finite element method (FEM), normal compliance, numerical approximation, Signorini condition, variational inequality, viscoplasticity.

Abstract

Contact phenomena abound, and play an important role in structural and mechanical engineering. Owing to their inherent complexity, they are modelled by highly nonlinear inequalities. Considerable progress has been achieved in modelling, variational analysis and numerical approximations of contact problems involving viscoelastic and viscoplastic materials. Moreover, it has led to several new types of variational inequalities. We present some recent results on the variational and numerical analysis of contact problems in viscoplasticity and some numerical examples of engineering applications.

1 Introduction

Contact phenomena among deformable bodies abound in industry and everyday life, and play an important role in structural and mechanical systems. The complicated surface structure, physics and chemistry involved in contact processes make it necessary to model them with highly nonlinear initial-boundary value problems. The famous Signorini problem was formulated in [23] as a model of unilateral frictionless contact between an elastic body and a rigid foundation. Mathematical analysis of this problem was first provided by Fichera [11]. Duvaut and Lions, in their monograph [6], systematically modelled and analyzed many important contact problems within the framework of the theory of variational inequalities. Numerical approximations of variational inequalities arising from contact problems were described in detail by Kikuchi and Oden [16], and Hlaváček *et al.* [14]. The mathematical, mechanical and numerical state of the art can be found in the proceedings Raous *et al.* [17], and in the special issue Shillor [22].

In earlier mathematical publications it was invariably assumed that the deformable bodies were linearly elastic. However, a number of recent publications is dedicated to the modelling, variational analysis and numerical approximations of contact problems involving viscoelastic and viscoplastic materials. Moreover, a variety of new and modified contact conditions were employed, reflecting the different settings and the nature of the problems. The settings studied were with unilateral or bilateral contact, with friction or frictionless. And in the cases of frictional contact, a number of different contact and friction conditions were employed.

Investigation of these problems led us to new variational inequalities, the well-posedness of which we established. Moreover, two types of numerical approximations were analyzed and error estimates were derived. These were the semi-discrete schemes, where only the spatial variables were discretized, and fully discrete schemes where both the time and the spatial variables were discretized. Here, we summarize our main recent results, and present a few numerical examples of engineering applications in viscoplastic materials.

In Section 2 we introduce notation and some preliminary material. In Section 3, we discuss several contact problems involving viscoplastic materials. We present the weak formulations, the well-posedness results and error

estimates for the numerical approximations. Because of space limitation, we only show results for the fully discrete schemes. In Section 4, we show some numerical examples.

2 Preliminaries

We consider mathematical models for quasistatic contact between a deformable body and a rigid or also deformable foundation. The physical setting is as follows. A deformable body occupies an open, bounded and connected set $\Omega \subset \mathbb{R}^d$, $d = 1$, 2 or 3. The boundary $\Gamma = \partial\Omega$ is assumed to be Lipschitz continuous and has the decomposition $\Gamma = \cup_{i=1}^3 \overline{\Gamma}_i$ into mutually disjoint, relatively open sets Γ_1, Γ_2 and Γ_3, with Lipschitz relative boundaries if $d = 3$. The set Γ_3 represents the potential contact surface, and we assume $\mathrm{meas}\,(\Gamma_1) > 0$. Since the boundary is Lipschitz continuous, the unit outward normal vector $\boldsymbol{\nu}$ exists a.e. on Γ.

We are interested in the evolution of the body's mechanical state over the time interval $[0, T]$ ($T > 0$). The body is clamped on Γ_1 and so the displacement field vanishes there. A surface traction of density $\boldsymbol{f}_2$ acts on Γ_2 and a volume force of density $\boldsymbol{f}_0$ acts in Ω, both depending on time. We assume that they change slowly in time so that the accelerations in the system are negligible, which means that the process is quasistatic.

We denote by $\mathbb{S}^d$ the space of second order symmetric tensors on $\mathbb{R}^d$, or equivalently, the space of symmetric matrices of order d. The inner products and the corresponding norms on $\mathbb{R}^d$ and $\mathbb{S}^d$ are

$$\begin{aligned} \boldsymbol{u} \cdot \boldsymbol{v} &= u_i v_i, \qquad \|\boldsymbol{v}\| = (\boldsymbol{v} \cdot \boldsymbol{v})^{1/2} \quad \forall\, \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d, \\ \boldsymbol{\sigma} \cdot \boldsymbol{\tau} &= \sigma_{ij}\tau_{ij}, \qquad \|\boldsymbol{\tau}\| = (\boldsymbol{\tau} \cdot \boldsymbol{\tau})^{1/2} \quad \forall\, \boldsymbol{\sigma}, \boldsymbol{\tau} \in \mathbb{S}^d. \end{aligned}$$

Here and below, $i, j = 1, 2, \ldots, d$, and the summation convention over repeated indices is adopted. Moreover, an index which follows a comma indicates a partial derivative. Let $\boldsymbol{\varepsilon}$ and Div be the *deformation* and *divergence* operators, respectively, defined by

$$\boldsymbol{\varepsilon}(\boldsymbol{u}) = (\varepsilon_{ij}(\boldsymbol{u})), \quad \varepsilon_{ij}(\boldsymbol{u}) = \frac{1}{2}\,(u_{i,j} + u_{j,i}), \quad \mathrm{Div}\,\boldsymbol{\sigma} = (\sigma_{ij,j}).$$

Denoting by $\boldsymbol{u}$ the displacement and $\boldsymbol{\sigma}$ the stress fields in the body, we have

$$\mathrm{Div}\,\boldsymbol{\sigma} + \boldsymbol{f}_0 = \boldsymbol{0} \quad \text{in } \Omega \times (0, T), \tag{1}$$

$$\boldsymbol{u} = \boldsymbol{0} \quad \text{on } \Gamma_1 \times (0, T), \tag{2}$$

$$\boldsymbol{\sigma}\boldsymbol{\nu} = \boldsymbol{f}_2 \quad \text{on } \Gamma_2 \times (0, T). \tag{3}$$

Here, (1) are the equilibrium equations, (2) and (3) are the displacement and the traction boundary conditions on Γ_1 and Γ_2, respectively. We need to supplement these relations with a constitutive law and a contact condition on $\Gamma_3 \times (0, T)$.

We need the following function spaces:

$$\begin{aligned} H &= \{\boldsymbol{u} = (u_i) \mid u_i \in L^2(\Omega)\}, \quad Q = \{\boldsymbol{\sigma} = (\sigma_{ij}) \mid \sigma_{ij} = \sigma_{ji} \in L^2(\Omega)\}, \\ H_1 &= \{\boldsymbol{u} = (u_i) \mid u_i \in H^1(\Omega)\}, \quad Q_1 = \{\boldsymbol{\sigma} \in Q \mid \sigma_{ij,j} \in H\}. \end{aligned}$$

These are real Hilbert spaces endowed with the inner products

$$\begin{aligned} (\boldsymbol{u}, \boldsymbol{v})_H &= \int_\Omega u_i v_i \, dx, \quad (\boldsymbol{\sigma}, \boldsymbol{\tau})_Q = \int_\Omega \sigma_{ij}\tau_{ij} \, dx, \\ (\boldsymbol{u}, \boldsymbol{v})_{H_1} &= (\boldsymbol{u}, \boldsymbol{v})_H + (\boldsymbol{\varepsilon}(\boldsymbol{u}), \boldsymbol{\varepsilon}(\boldsymbol{v}))_Q, \quad (\boldsymbol{\sigma}, \boldsymbol{\tau})_{Q_1} = \boldsymbol{\sigma}, \boldsymbol{\tau})_Q + (\mathrm{Div}\,\boldsymbol{\sigma}, \mathrm{Div}\,\boldsymbol{\tau})_H, \end{aligned}$$

and the associated norms are denoted by $\|\cdot\|_H$, $\|\cdot\|_Q$, $\|\cdot\|_{H_1}$ and $\|\cdot\|_{Q_1}$.

Everywhere in this paper, unless stated otherwise, V stands for the space $V = \{\boldsymbol{v} \in H^1(\Omega)^d \mid \boldsymbol{v} = \boldsymbol{0} \text{ on } \Gamma_1\}$ equipped with the inner product

$$(\boldsymbol{u}, \boldsymbol{v})_V = (\boldsymbol{\varepsilon}(\boldsymbol{u}), \boldsymbol{\varepsilon}(\boldsymbol{v}))_Q \quad \forall\, \boldsymbol{u},\, \boldsymbol{v} \in V \tag{4}$$

and the associated norm $\|\cdot\|_V$. Since $\mathrm{meas}(\Gamma_1) > 0$, it follows from Korn's inequality that $\|\cdot\|_{H^1(\Omega)^d}$ and $\|\cdot\|_V$ are equivalent norms on V.

For an element $\boldsymbol{v} \in H_1$, we also denote by $\boldsymbol{v}$ its trace $\gamma\boldsymbol{v}$ on Γ; v_ν and $\boldsymbol{v}_\tau$ denote the *normal* and *tangential* components of $\boldsymbol{v}$ on Γ given by $v_\nu = \boldsymbol{v}\cdot\boldsymbol{\nu}, \quad \boldsymbol{v}_\tau = \boldsymbol{v} - v_\nu\boldsymbol{\nu}$. For an element $\boldsymbol{\sigma} \in Q_1$, we denote by $\boldsymbol{\sigma\nu}$ its trace on Γ. If $\boldsymbol{\sigma}$ is a smooth function (e.g. continuously differentiable on $\overline{\Omega}$), then

$$(\boldsymbol{\sigma}, \boldsymbol{\varepsilon}(\boldsymbol{v}))_Q + (\mathrm{Div}\,\boldsymbol{\sigma}, \boldsymbol{v})_H = \int_\Gamma \boldsymbol{\sigma\nu}\cdot\boldsymbol{v}\,da$$

for all $\boldsymbol{v} \in H_1$, where da is the surface measure. In this case the *normal* and *tangential* components of $\boldsymbol{\sigma}$ are given by $\sigma_\nu = (\boldsymbol{\sigma\nu})\cdot\boldsymbol{\nu}, \quad \boldsymbol{\sigma}_\tau = \boldsymbol{\sigma\nu} - \sigma_\nu\boldsymbol{\nu}$.

Given a real normed space $(X, \|\cdot\|_X)$ we denote by $C([0,T];X)$ and $C^1([0,T];X)$ the spaces of continuous and continuously differentiable functions from $[0,T]$ to X with the respective norms

$$\|x\|_{C([0,T];X)} = \max_{t\in[0,T]} \|x(t)\|_X, \quad \|x\|_{C^1([0,T];X)} = \max_{t\in[0,T]} \|x(t)\|_X + \max_{t\in[0,T]} \|\dot{x}(t)\|_X.$$

Here and below, a dot above a variable represents its derivative with respect to time. For an integer $k \geq 0$, and $p \in [1,\infty]$, $W^{k,p}(0,T;X)$ is the Sobolev space of the vector-valued functions x such that

$$\|x\|_{W^{k,p}(0,T;X)} = \sum_{j=0}^{k} \|x^{(j)}\|_{L^p(0,T;X)} < \infty.$$

In our numerical approximations of the problems, we use the finite element method (FEM) for spatial discretization, and finite differences for the temporal derivative. We now describe briefly a finite dimensional space H_1^h, which approximates H_1, via the FEM. The details can be found in, e.g., [5]. For the sake of simplicity, we assume that $\overline{\Omega}$ is a polygon or polyhedron. Then $\overline{\Gamma}_3 = \cup_{i=1}^I \overline{\Gamma}_{3,i}$, and each piece $\overline{\Gamma}_{3,i}$ is represented by an affine function. Let $\mathcal{T}^h$ be a regular finite element partition of $\overline{\Omega}$ in such a way that if a side of an element lies on the boundary, the side belongs entirely to one of the subsets $\overline{\Gamma}_1$, $\overline{\Gamma}_2$ and $\overline{\Gamma}_{3,i}$, $1 \leq i \leq I$. Let h be the maximal diameter of the elements. We define $H_1^h \subset H_1$ to be the finite element space consisting of piecewise linear functions, corresponding to the partition $\mathcal{T}^h$. If the solution $\boldsymbol{u}$ is known to have higher regularity, we may use higher order elements; our error analysis can be easily extended to such cases.

We employ the partition of the time interval $[0,T] : 0 = t_0 < t_1 < \cdots < t_N = T$. We denote the step-size by $k_n = t_n - t_{n-1}$, for $n = 1,\ldots,N$, and let $k = \max_n k_n$ be the maximal step-size. For a continuous function $w(t)$, we let $w_n = w(t_n)$. Given a sequence $\{w_n\}_{n=0}^N$, for $n = 1,\ldots,N$, we denote $\Delta w_n = w_n - w_{n-1}$, and let $\delta w_n = \Delta w_n / k_n$ be the corresponding divided difference, where no summation is implied over the index n.

Everywhere below, the symbol c represents a positive constant which may change its value from place to place and may depend on the input data, but it is independent of discretization parameters h and k.

3 Contact problems in viscoplasticity

We use the rate-type viscoplastic constitutive law

$$\dot{\boldsymbol{\sigma}} = \mathcal{E}\boldsymbol{\varepsilon}(\dot{\boldsymbol{u}}) + G(\boldsymbol{\sigma}, \boldsymbol{\varepsilon}(\boldsymbol{u})), \tag{5}$$

where $\mathcal{E}$ and G are material constitutive functions. The function $\mathcal{E}$ is assumed to be linear while G is nonlinear.

The Perzyna laws is an example of such elastic-viscoplastic constitutive law,

$$\dot{\boldsymbol{\varepsilon}} = \mathcal{E}^{-1}\dot{\boldsymbol{\sigma}} + \frac{1}{\mu^*}(\boldsymbol{\sigma} - P_K\boldsymbol{\sigma}),$$

in which $\mu^* > 0$ is the viscosity constant, K is a nonempty, closed, convex set in the space of symmetric tensors and P_K is the projection mapping on K. Note that G does not depend on $\boldsymbol{\varepsilon}$.

Rate-type viscoplastic models of the form (5) have been used to describe the behavior of rubbers, metals, pastes, rocks, etc. Models of mechanical problems of this form may be found in [2] (see also references therein). Existence and uniqueness results for initial–boundary value problems involving (5) were obtained in [15] for displacements-tractions conditions.

We assume in the sequel that $\mathcal{E} = (\mathcal{E}_{ijkl})$ and $G : \Omega \times \mathbb{S}^d \times \mathbb{S}^d \to \mathbb{S}^d$ satisfy the assumptions:

$$\left.\begin{array}{l} \text{(a) } \mathcal{E}_{ijkl} \in L^\infty(\Omega),\ 1 \le i,j,k,l \le d. \\ \text{(b) } \mathcal{E}\boldsymbol{\sigma} \cdot \boldsymbol{\tau} = \boldsymbol{\sigma} \cdot \mathcal{E}\boldsymbol{\tau},\ \forall\, \boldsymbol{\sigma}, \boldsymbol{\tau} \in S_d \quad \text{a.e. in } \Omega. \\ \text{(c) There exists an } \alpha_0 > 0 \text{ such that } \mathcal{E}\boldsymbol{\tau} \cdot \boldsymbol{\tau} \ge \alpha_0 |\boldsymbol{\tau}|^2 \quad \forall\, \boldsymbol{\tau} \in S_d, \text{ a.e. in } \Omega. \end{array}\right\} \tag{6}$$

$$\left.\begin{array}{l} \text{(a) There exists an } \mathcal{L} > 0 \text{ such that} \\ \|G(\boldsymbol{x}, \boldsymbol{\sigma}_1, \boldsymbol{\varepsilon}_1) - G(\boldsymbol{x}, \boldsymbol{\sigma}_2, \boldsymbol{\varepsilon}_2)\| \le \mathcal{L}\left(\|\boldsymbol{\sigma}_1 - \boldsymbol{\sigma}_2\| + \|\boldsymbol{\varepsilon}_1 - \boldsymbol{\varepsilon}_2\|\right) \quad \forall\, \boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2, \boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2 \in \mathbb{S}^d, \text{ a.e. in } \Omega. \\ \text{(b) For any } \boldsymbol{\sigma}, \boldsymbol{\varepsilon} \in \mathbb{S}^d, \text{ the mapping } \boldsymbol{x} \mapsto G(\boldsymbol{x}, \boldsymbol{\sigma}, \boldsymbol{\varepsilon}) \text{ is measurable.} \\ \text{(c) The mapping } \boldsymbol{x} \mapsto G(\boldsymbol{x}, \boldsymbol{0}, \boldsymbol{0}) \text{ belongs to } Q. \end{array}\right\} \tag{7}$$

Forces and tractions are assumed to satisfy:

$$\boldsymbol{f}_0 \in W^{1,\infty}(0,T;H), \quad \boldsymbol{f}_2 \in W^{1,\infty}(0,T;L^2(\Gamma_2)^d) \tag{8}$$

and we denote by $\boldsymbol{f} \in W^{1,\infty}(0,T;V)$ the unique element given by

$$(\boldsymbol{f}(t), \boldsymbol{v})_V = (\boldsymbol{f}_0(t), \boldsymbol{v})_H + (\boldsymbol{f}_2(t), \boldsymbol{v})_{L^2(\Gamma_2)^d} \quad \forall\, \boldsymbol{v} \in V, \forall\, t \in [0,T]. \tag{9}$$

We present now a number of contact problems involving viscoplastic materials of the type (5).

3.1 The Signorini problem

We assume the contact without friction and there is no penetration between the body and the foundation. The classical formulation of the problem is the following:

Find a displacement field $\boldsymbol{u} : \Omega \times [0,T] \to \mathbb{R}^d$ and a stress field $\boldsymbol{\sigma} : \Omega \times [0,T] \to \mathbb{S}^d$ satisfying (1)–(3), and

$$\dot{\boldsymbol{\sigma}} = \mathcal{E}\boldsymbol{\varepsilon}(\dot{\boldsymbol{u}}) + G(\boldsymbol{\sigma}, \boldsymbol{\varepsilon}(\boldsymbol{u})) \quad \text{in } \Omega \times (0,T), \tag{10}$$

$$u_\nu \le 0,\ \sigma_\nu \le 0,\ \sigma_\nu u_\nu = 0 \quad \text{on } \Gamma_3 \times (0,T), \tag{11}$$

$$\boldsymbol{\sigma}_\tau = \boldsymbol{0} \quad \text{on } \Gamma_3 \times (0,T), \tag{12}$$

$$\boldsymbol{u}(0) = \boldsymbol{u}_0, \quad \boldsymbol{\sigma}(0) = \boldsymbol{\sigma}_0 \quad \text{in } \Omega. \tag{13}$$

Here, $\boldsymbol{u}_0$ and $\boldsymbol{\sigma}_0$ are given initial data, expressions (11)and (12)are the nonpenetration (Signorini) and no friction conditions, respectively . Let $U = \{\boldsymbol{v} \in V \mid v_\nu \le 0 \text{ on } \Gamma_3\}$, where V is defined in Section 2. Assume, for the initial data

$$\boldsymbol{u}_0 \in U, \quad \boldsymbol{\sigma}_0 \in Q, \tag{14}$$

$$(\boldsymbol{\sigma}_0, \boldsymbol{v} - \boldsymbol{\varepsilon}(\boldsymbol{u}_0))_Q \ge (\boldsymbol{f}(0), \boldsymbol{v} - \boldsymbol{u}_0)_V, \forall \boldsymbol{v} \in U \tag{15}$$

The weak formulation for the contact problem is:

Problem 3.1 *Find a displacement $\boldsymbol{u} : [0,T] \to U$ and the stress tensor $\boldsymbol{\sigma} : [0,T] \to Q$ such that $\boldsymbol{u}(0) = \boldsymbol{u}_0$, $\boldsymbol{\sigma}(0) = \boldsymbol{\sigma}_0$ and, for a.e. $t \in (0,T)$,*

$$\dot{\boldsymbol{\sigma}}(t) = \mathcal{E}\boldsymbol{\varepsilon}(\dot{\boldsymbol{u}}(t)) + G(\boldsymbol{\sigma}(t), \boldsymbol{\varepsilon}(\boldsymbol{u}(t))),$$
$$(\boldsymbol{\sigma}(t), \boldsymbol{\varepsilon}(\boldsymbol{v}) - \boldsymbol{\varepsilon}(\boldsymbol{u}(t)))_Q \ge (\boldsymbol{f}(t), \boldsymbol{v} - \boldsymbol{u}(t))_V \ \ \forall\, \boldsymbol{v} \in U.$$

The following result has been established in [24].

Theorem 3.2 *Assume that* (6), (7), (8), (14) *and* (15) *hold. Then the problem 3.1 has a unique solution $\boldsymbol{u} \in W^{1,\infty}(0,T;U)$, $\boldsymbol{\sigma} \in W^{1,\infty}(0,T;Q_1)$.*

For numerical approximations, let $V^h \subset V$ be a finite dimensional subspace of V and define $U^h = \{\boldsymbol{v}^h \in V^h \mid v^h_\nu \le 0 \text{ on } \Gamma_3\}$. Let $Q^h \subset Q$ be a finite dimensional subspace of Q such that $\boldsymbol{\varepsilon}(V^h) \subset Q^h$. Let $\mathcal{P}_{Q^h} : Q \to Q^h$ be the orthogonal projection defined by $(\mathcal{P}_{Q^h}\boldsymbol{q}, \boldsymbol{q}^h)_Q = (\boldsymbol{q}, \boldsymbol{q}^h)_Q \quad \forall\, \boldsymbol{q} \in Q,\ \boldsymbol{q}^h \in Q^h$. Then a fully discrete approximation of problem 3.1 is:

Problem 3.3 *Given $\boldsymbol{u}_0^h \in U^h$ and $\boldsymbol{\sigma}_0^h \in Q^h$ find $\boldsymbol{u}^{hk} = \{\boldsymbol{u}_n^{hk}\}_{n=1}^N \subset U^h$ and $\boldsymbol{\sigma}^{hk} = \{\boldsymbol{\sigma}_n^{hk}\}_{n=1}^N$ in Q^h such that $\boldsymbol{u}_0^{hk} = \boldsymbol{u}_0^h$, $\boldsymbol{\sigma}_0^{hk} = \boldsymbol{\sigma}_0^h$, and, for $n = 1, \ldots, N$,*

$$\delta\boldsymbol{\sigma}_n^{hk} = \mathcal{P}_{Q^h}\mathcal{E}\delta\boldsymbol{\varepsilon}(\boldsymbol{u}_n^{hk}) + \mathcal{P}_{Q^h}G(\boldsymbol{\varepsilon}(\boldsymbol{u}_n^{hk}), \boldsymbol{\sigma}_n^{hk}),$$
$$(\boldsymbol{\sigma}_n^{hk}, \boldsymbol{\varepsilon}(\boldsymbol{v}^h - \boldsymbol{u}_n^{hk}))_Q \geq (\boldsymbol{f}_n, \boldsymbol{v}^h - \boldsymbol{u}_n^{hk})_V \; \forall\, \boldsymbol{v}^h \in U^h.$$

Problem 3.3 has a unique solution for k small enough. We obtain the following error estimates by modifying the results in [3](see also [7]). Remark that they are satisfied when V^h is the space of piecewise linear polynomials and Q^h the space of piecewise constant functions.

Theorem 3.4 *Assume that the conditions in Theorem 3.2 hold and also the following ones:*

- $\|\boldsymbol{u}_0 - \boldsymbol{u}_0^h\|_V \leq ch, \quad \|\boldsymbol{\sigma}_0 - \boldsymbol{\sigma}_0^h\|_Q \leq ch,$
- $[H^2(\Omega)]^d \cap U$ *is dense in* U*:* $\inf_{\boldsymbol{v}^h \in U^h} \|\boldsymbol{v} - \boldsymbol{v}^h\|_V \leq c(\boldsymbol{v})h, \quad \forall \boldsymbol{v} \in [H^2(\Omega)]^d \cap U,$
- $\|(I_Q - \mathcal{P}_{Q^h})\boldsymbol{\tau}\|_Q \leq ch, \quad \forall \boldsymbol{\tau} \in Q,$
- $\boldsymbol{u} \in L^\infty(0, T; [H^2(\Omega)]^d),$

then

$$\max_{1\leq n\leq N} \left(\|\boldsymbol{u}_n - \boldsymbol{u}_n^{hk}\|_V + \|\boldsymbol{\sigma}_n - \boldsymbol{\sigma}_n^{hk}\|_Q\right)$$
$$\leq ch^{1/2}\|\boldsymbol{u}\|_{L^\infty(0,T;[H^2(\Omega)]^d)} + c\,k\,\left(\|\dot{\boldsymbol{u}}\|_{L^\infty(0,T;V)} + \|\dot{\boldsymbol{\sigma}}\|_{L^\infty(0,T;Q)}\right).$$

If we further assume

- $u_\nu \in L^\infty(0, T; H^2(\Gamma_3)), \quad \sigma_\nu \in L^\infty(0, T; L^2(\Gamma_3))$
- $\inf_{\boldsymbol{v}^h \in U^h} \left[\|\boldsymbol{v} - \boldsymbol{v}^h\|_V + \|v_\nu - v_\nu^h\|_{L^2(\Gamma_3)}^{1/2}\right] \leq c(\boldsymbol{v})h, \; \forall \boldsymbol{v} \in [H^2(\Omega)]^d \cap U,$

then we have an optimal order error estimate

$$\max_{1\leq n\leq N} \left(\|\boldsymbol{u}_n - \boldsymbol{u}_n^{hk}\|_V + \|\boldsymbol{\sigma}_n - \boldsymbol{\sigma}_n^{hk}\|_Q\right)$$
$$\leq ch\left(\|\boldsymbol{u}\|_{L^\infty(0,T;[H^2(\Omega)]^d)} + \|u_\nu\|_{L^\infty(0,T;H^2(\Gamma_3))}^{1/2}\right) + c\,k\,\left(\|\dot{\boldsymbol{u}}\|_{L^\infty(0,T;V)} + \|\dot{\boldsymbol{\sigma}}\|_{L^\infty(0,T;Q)}\right). \quad (16)$$

3.2 Frictionless contact problems with normal compliance

We consider frictionless contact with a deformable foundation which we model by

$$-\sigma_\nu = r^*\,(u_\nu - g)_+^\alpha, \quad \boldsymbol{\sigma}_\tau = \boldsymbol{0} \quad \text{on } \Gamma_3 \times (0, T). \quad (17)$$

Here $\alpha \in (0, 1]$, g is the initial gap between the elastic-viscoplastic body and the foundation and $1/r^*$ may be interpreted as the *coefficient of deformability* of the foundation. We assume

$$g \in L^2(\Gamma_3), \quad g \geq 0, \quad r^* \in L^\infty(\Gamma_3), \quad r^* > 0 \quad \text{a. e. on } \Gamma_3. \quad (18)$$

Condition (17) is the normal compliance condition. The expression $u_\nu - g$, when positive, represents the penetration of the body into the foundation. Signorini's nonpenetration condition is obtained from (17) when $r^* \to \infty$, i.e. when the coefficient of deformability of the foundation tends to zero. Then, the classical formulation of the problem is to find a displacement field $\boldsymbol{u} : \Omega \times [0, T] \to \mathbb{R}^d$ and a stress field $\boldsymbol{\sigma} : \Omega \times [0, T] \to \mathbb{S}^d$ satisfying (1)–(3), (10), (13) and (17).

Let

$$j(\boldsymbol{u}, \boldsymbol{v}) = \int_{\Gamma_3} r^*(u_\nu - g)_+^\alpha v_\nu, \; \forall \boldsymbol{u}, \boldsymbol{v} \in V. \quad (19)$$

Assume for the initial data,

$$\boldsymbol{u}_0 \in V, \quad \boldsymbol{\sigma}_0 \in Q, \quad (\boldsymbol{\sigma}_0, \boldsymbol{\varepsilon}(\boldsymbol{v}))_Q + j(\boldsymbol{u}_0, \boldsymbol{v}) = (\boldsymbol{f}(0), \boldsymbol{v})_V \quad \forall\, \boldsymbol{v} \in V. \quad (20)$$

The weak formulation for the contact problem is:

Problem 3.5 *Find a displacement* $\boldsymbol{u} : [0,T] \to V$ *and a stress tensor* $\boldsymbol{\sigma} : [0,T] \to Q$ *satisfying* $\boldsymbol{u}(0) = \boldsymbol{u}_0$, $\boldsymbol{\sigma}(0) = \boldsymbol{\sigma}_0$, *and, for a.e.* $t \in (0,T)$,

$$\dot{\boldsymbol{\sigma}}(t) = \mathcal{E}\boldsymbol{\varepsilon}(\dot{\boldsymbol{u}}(t)) + G(\boldsymbol{\sigma}(t), \boldsymbol{\varepsilon}(\boldsymbol{u}(t))),$$
$$(\boldsymbol{\sigma}(t), \boldsymbol{\varepsilon}(\boldsymbol{v}))_Q + j(\boldsymbol{u}(t), \boldsymbol{v}) = (\boldsymbol{f}(t), \boldsymbol{v})_V \quad \forall\, \boldsymbol{v} \in V.$$

Well-posedness of the problem 3.5 has been studied in [9].

Theorem 3.6 *Assume* (6), (7), (8), (18)–(20). *Then the problem 3.5 has a unique solution* $\boldsymbol{u} \in W^{1,\infty}(0,T;V)$, $\boldsymbol{\sigma} \in W^{1,\infty}(0,T;Q_1)$.

For numerical approximations, let $V^h \subset V$ and $Q^h \subset Q$ be finite-dimensional spaces. We assume that these spaces satisfy $\boldsymbol{\varepsilon}(V^h) \subset Q^h$. This assumption is very natural and holds for finite element approximations when the polynomial degree for the space V^h is at most one higher than that for the space Q^h. Then a fully discrete approximation to the problem 3.5 is:

Problem 3.7 *Given* $\boldsymbol{u}_0^h \in V^h$ *and* $\boldsymbol{\sigma}_0^h \in Q^h$, *find* $\boldsymbol{u}^{hk} = \{\boldsymbol{u}_n^{hk}\}_{n=1}^N \subset V^h$ *and* $\boldsymbol{\sigma}^{hk} = \{\boldsymbol{\sigma}_n^{hk}\}_{n=1}^N \subset Q^h$ *such that* $\boldsymbol{u}_0^{hk} = \boldsymbol{u}_0^h$, $\boldsymbol{\sigma}_0^{hk} = \boldsymbol{\sigma}_0^h$, *and, for* $n = 1, \ldots, N$,

$$\delta\boldsymbol{\sigma}_n^{hk} = \mathcal{P}_{Q^h}\mathcal{E}\delta\boldsymbol{\varepsilon}(\boldsymbol{u}_n^{hk}) + \mathcal{P}_{Q^h}G(\boldsymbol{\varepsilon}(\boldsymbol{u}_n^{hk}), \boldsymbol{\sigma}_n^{hk}),$$
$$(\boldsymbol{\sigma}_n^{hk}, \boldsymbol{\varepsilon}(\boldsymbol{v}^h))_Q + j(\boldsymbol{u}_n^{hk}, \boldsymbol{v}^h) = (\boldsymbol{f}_n, \boldsymbol{v}^h)_V \quad \forall\, \boldsymbol{v}^h \in V^h.$$

Problem 3.7 has a unique solution for k small enough and we have the following error estimates(see [9])which can be applied in the habitual case with V^h the space of piecewise linear polynomials and Q^h the space of piecewise constant functions.

Theorem 3.8 *Assume that the conditions in Theorem 3.6 hold and also the following ones:*

- $\|\boldsymbol{u}_0 - \boldsymbol{u}_0^h\|_V \le ch, \quad \|\boldsymbol{\sigma}_0 - \boldsymbol{\sigma}_0^h\|_Q \le ch$,
- $[H^2(\Omega)]^d \cap V$ *is dense in* V: $\inf_{\boldsymbol{v}^h \in V^h} \|\boldsymbol{v} - \boldsymbol{v}^h\|_V \le c(\boldsymbol{v})h, \quad \forall \boldsymbol{v} \in [H^2(\Omega)]^d \cap V$,
- $\|(I_Q - \mathcal{P}_{Q^h})\boldsymbol{\tau}\|_Q \le ch, \quad \forall \boldsymbol{\tau} \in Q$,
- $\boldsymbol{u} \in L^\infty(0,T;[H^2(\Omega)]^d)$,

then

$$\max_{1\le n\le N} \left(\|\boldsymbol{u}_n - \boldsymbol{u}_n^{hk}\|_V + \|\boldsymbol{\sigma}_n - \boldsymbol{\sigma}_n^{hk}\|_Q\right)$$
$$\le ch\|\boldsymbol{u}\|_{L^\infty(0,T;[H^2(\Omega)]^d)} + c\,k\,\left(\|\dot{\boldsymbol{u}}\|_{L^\infty(0,T;V)} + \|\dot{\boldsymbol{\sigma}}\|_{L^\infty(0,T;Q)}\right). \tag{21}$$

Variational and numerical analysis of a quasistatic frictionless contact problem for viscoplastic materials with a general normal compliance contact condition have been obtained in [9].

3.3 Frictionless contact between two viscoplastic bodies

We consider two elastic-viscoplastic bodies occupying two bounded domains Ω^1 and Ω^2 of $\mathbb{R}^d$ ($d \le 3$). We use the superscript m to indicate that the variable is related to Ω^m, where here and below $m = 1, 2$. For each domain Ω^m, we assume its boundary Γ^m is Lipschitz continuous, and is partitioned into three disjoint measurable parts Γ_1^m, Γ_2^m and Γ_3^m, with $\mathrm{meas}\,(\Gamma_1^m) > 0$. The unit outward normal to Γ^m, is denoted by $\boldsymbol{\nu}^m = (\nu_i^m)$. We are interested in the evolution of the contact process over $[0,T]$, ($T > 0$). The bodies are clamped on $\Gamma_1^m \times (0,T)$, volume forces of density $\boldsymbol{f}_0^m$ act on $\Omega^m \times (0,T)$ and surface tractions of density $\boldsymbol{f}_2^m$ act on $\Gamma_2^m \times (0,T)$. The two bodies are in contact along the common part $\Gamma_3^1 = \Gamma_3^2$, denoted by Γ_3 below. The contact is frictionless and we model it by the Signorini condition on $\Gamma_3^m \times (0,T)$ with vanishing gap function. Finally, we assume that the process is quasistatic and use (5) as constitutive law. The mechanical problem we study is formulated as follows:

Find displacement fields $\boldsymbol{u}^m = (u_i^m) : \Omega^m \times [0,T] \to \mathbb{R}^d$ and stress fields $\boldsymbol{\sigma}^m = (\sigma_{ij}^m) : \Omega^m \times [0,T] \to \mathbb{S}^d$, $m = 1, 2$, which satisfy

$$\begin{aligned}
\dot{\boldsymbol{\sigma}}^m &= \mathcal{E}^m \boldsymbol{\varepsilon}(\dot{\boldsymbol{u}}^m) + G^m(\boldsymbol{\sigma}^m, \boldsymbol{\varepsilon}(\boldsymbol{u}^m)) \quad \text{in } \Omega^m \times (0,T),\\
\operatorname{Div} \boldsymbol{\sigma}^m + \boldsymbol{f}_0^m &= \boldsymbol{0} \quad \text{in } \Omega^m \times (0,T),\\
\boldsymbol{u}^m &= \boldsymbol{0} \quad \text{on } \Gamma_1^m \times (0,T),\\
\boldsymbol{\sigma}^m \boldsymbol{\nu}^m &= \boldsymbol{f}_2^m \quad \text{on } \Gamma_2^m \times (0,T),\\
u_\nu^1 + u_\nu^2 \le 0,\ \sigma_\nu^1 &= \sigma_\nu^2 \le 0,\\
\sigma_\nu^1 (u_\nu^1 + u_\nu^2) = 0,\ \boldsymbol{\sigma}_\tau^m &= \boldsymbol{0} \quad \text{on } \Gamma_3 \times (0,T),
\end{aligned}$$

and the initial conditions

$$\boldsymbol{u}^m(0) = \boldsymbol{u}_0^m,\ \boldsymbol{\sigma}^m(0) = \boldsymbol{\sigma}_0^m \quad \text{in } \Omega^m.$$

We introduce the spaces

$$\begin{aligned}
V^m &= \{\boldsymbol{v} = (v_i) \mid v_i \in H^1(\Omega^m),\ v_i = 0 \text{ on } \Gamma_1^m,\ 1 \le i \le d\},\\
Q^m &= \{\boldsymbol{\tau} = (\tau_{ij}) \mid \tau_{ij} \in L^2(\Omega^m),\ 1 \le i,j \le d\},\\
Q_1^m &= \{\boldsymbol{\tau} \in Q^m \mid \operatorname{Div} \boldsymbol{\tau} \in L^2(\Omega^m)^d\}.
\end{aligned}$$

These are Hilbert spaces with their canonical inner products. Since $\operatorname{meas}(\Gamma_1^m) > 0$, by Korn's inequality, $\|\boldsymbol{\varepsilon}(\boldsymbol{v})\|_{Q^m}$ is a norm on $H^1(\Omega^m)^d$ and is equivalent to $\|\boldsymbol{v}\|_{H^1(\Omega^m)^d}$.

We make the following assumptions on : $\mathcal{E}^m = (\mathcal{E}_{ijkl}^m)$ and $G^m : \Omega^m \times S_d \times \mathbb{S}^d \to \mathbb{S}^d$:

$$\left.\begin{array}{l}
\text{(a) } \mathcal{E}_{ijkl}^m \in L^\infty(\Omega^m),\ 1 \le i,j,k,l \le d;\\
\text{(b) } \mathcal{E}^m \boldsymbol{\sigma} \cdot \boldsymbol{\tau} = \boldsymbol{\sigma} \cdot \mathcal{E}^m \boldsymbol{\tau}\ \forall\, \boldsymbol{\sigma}, \boldsymbol{\tau} \in \mathbb{S}^d, \text{ a.e. in } \Omega^m;\\
\text{(c) There exists an } \alpha^m > 0 \text{ such that } \mathcal{E}^m \boldsymbol{\tau} \cdot \boldsymbol{\tau} \ge \alpha^m |\boldsymbol{\tau}|^2\ \forall\, \boldsymbol{\tau} \in \mathbb{S}^d, \text{ a.e. in } \Omega^m.
\end{array}\right\} \tag{22}$$

$$\left.\begin{array}{l}
\text{(a) There exists an } L^m > 0 \text{ such that } \|G^m(\boldsymbol{x}, \boldsymbol{\sigma}_1, \boldsymbol{\varepsilon}_1) - G^m(\boldsymbol{x}, \boldsymbol{\sigma}_2, \boldsymbol{\varepsilon}_2)\|\\
\qquad \le L^m \left(\|\boldsymbol{\sigma}_1 - \boldsymbol{\sigma}_2\| + \|\boldsymbol{\varepsilon}_1 - \boldsymbol{\varepsilon}_2\|\right) \forall\, \boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2, \boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2 \in \mathbb{S}^d, \text{ a.e. in } \Omega^m;\\
\text{(b) } \forall\, \boldsymbol{\sigma}, \boldsymbol{\varepsilon} \in \mathbb{S}^d, \text{ the mapping } \boldsymbol{x} \mapsto G^m(\boldsymbol{x}, \boldsymbol{\sigma}, \boldsymbol{\varepsilon}) \text{ is measurable;}\\
\text{(c) The mapping } \boldsymbol{x} \mapsto G^m(\boldsymbol{x}, \boldsymbol{0}, \boldsymbol{0}) \text{belongs to } Q^m.
\end{array}\right\} \tag{23}$$

The force densities satisfy:

$$\boldsymbol{f}_0^m \in W^{1,\infty}(0,T; L^2(\Omega^m)^d), \quad \boldsymbol{f}_2^m \in W^{1,\infty}(0,T; L^2(\Gamma_2^m)^d). \tag{24}$$

We define the product spaces $V = V^1 \times V^2$, $Q = Q^1 \times Q^2$ and $Q_1 = Q_1^1 \times Q_1^2$. These are all Hilbert spaces endowed with the canonical inner products $(\cdot,\cdot)_V$, $(\cdot,\cdot)_Q$ and $(\cdot,\cdot)_{Q_1}$, respectively. The associated norms are $\|\cdot\|_V$, $\|\cdot\|_Q$ and $\|\cdot\|_{Q_1}$, respectively. Moreover, $(\cdot,\cdot)_V$ is the inner product on V.

Let $\boldsymbol{f}(t)$ denote the element of V, for $t \in [0,T]$, given by

$$(\boldsymbol{f}(t), \boldsymbol{v})_V = (\boldsymbol{f}_0^1(t), \boldsymbol{v}^1)_{L^2(\Omega^1)^d} + (\boldsymbol{f}_0^2(t), \boldsymbol{v}^2)_{L^2(\Omega^2)^d} + (\boldsymbol{f}_2^1(t), \boldsymbol{v}^1)_{L^2(\Gamma_2^1)^d} + (\boldsymbol{f}_2^2(t), \boldsymbol{v}^2)_{L^2(\Gamma_2^2)^d},$$

for all $\boldsymbol{v} = (\boldsymbol{v}^1, \boldsymbol{v}^2) \in V$. We define the set U of admissible displacement fields by

$$U = \{\boldsymbol{v} = (\boldsymbol{v}^1, \boldsymbol{v}^2) \in V \mid v_\nu^1 + v_\nu^2 \le 0 \text{ on } \Gamma_3\}, \tag{25}$$

and we suppose that

$$\boldsymbol{u}_0 = (\boldsymbol{u}_0^1, \boldsymbol{u}_0^2) \in U, \quad \boldsymbol{\sigma}_0 = (\boldsymbol{\sigma}_0^1, \boldsymbol{\sigma}_0^2) \in Q, \quad (\boldsymbol{\sigma}_0, \boldsymbol{\varepsilon}(\boldsymbol{v} - \boldsymbol{u}_0))_Q \ge (\boldsymbol{f}(0), \boldsymbol{v} - \boldsymbol{u}_0)_V. \tag{26}$$

Finally, we use the notation $\boldsymbol{\varepsilon}(\boldsymbol{v}) = (\boldsymbol{\varepsilon}(\boldsymbol{v}^1), \boldsymbol{\varepsilon}(\boldsymbol{v}^2))$ for $\boldsymbol{v} = (\boldsymbol{v}^1, \boldsymbol{v}^2) \in V$ and $\mathcal{E}\boldsymbol{\varepsilon} = (\mathcal{E}^1\boldsymbol{\varepsilon}^1, \mathcal{E}^2\boldsymbol{\varepsilon}^2)$, $G(\boldsymbol{\sigma}, \boldsymbol{\varepsilon}) = (G(\boldsymbol{\sigma}^1, \boldsymbol{\varepsilon}^1), G(\boldsymbol{\sigma}^2, \boldsymbol{\varepsilon}^2))$ for $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}^1, \boldsymbol{\varepsilon}^2) \in Q$ and $\boldsymbol{\sigma} = (\boldsymbol{\sigma}^1, \boldsymbol{\sigma}^2) \in Q$.

The weak formulation of the contact problem is:

Problem 3.9 *Find a displacement field* $\boldsymbol{u} : [0,T] \to U$ *and a stress field* $\boldsymbol{\sigma} : [0,T] \to Q_1$ *such that* $\boldsymbol{u}(0) = \boldsymbol{u}_0$, $\boldsymbol{\sigma}(0) = \boldsymbol{\sigma}_0$, *and, for a.e.* $t \in (0,T)$,

$$\dot{\boldsymbol{\sigma}}(t) = \mathcal{E}\boldsymbol{\varepsilon}(\dot{\boldsymbol{u}}(t)) + G(\boldsymbol{\sigma}(t), \boldsymbol{\varepsilon}(\boldsymbol{u}(t))),$$
$$(\boldsymbol{\sigma}(t), \boldsymbol{\varepsilon}(\boldsymbol{v} - \boldsymbol{u}(t)))_Q \geq (\boldsymbol{f}(t), \boldsymbol{v} - \boldsymbol{u}(t))_V \quad \forall\, \boldsymbol{v} \in U.$$

The well-posedness of the problem 3.9 has been established in [18]; the main existence and uniqueness result is the following.

Theorem 3.10 *Under the assumptions* (22), (23), (24), *and* (26), *Problem 3.9 has a unique solution* $(\boldsymbol{u}, \boldsymbol{\sigma}) \in W^{1,\infty}(0,T; U \times Q_1)$.

We turn to numerical approximations. Let $\mathcal{T}^h$ be a regular FEM partition of the domain Ω in such a way that if a side of an element lies on the boundary, then the side is entirely on one of the subsets $\overline{\Gamma}_1^m$, $\overline{\Gamma}_2^m$ and $\overline{\Gamma}_3$. We choose a finite element space $V^h \subset V$ for the approximation of $\boldsymbol{u}$, and another FEM space Q^h such that $\boldsymbol{\varepsilon}(V^h) \subset Q^h$, for the approximation of $\boldsymbol{\sigma}$. Then, we define the discrete admissible set

$$U^h = \{\boldsymbol{v}^h = (\boldsymbol{v}^{1,h}, \boldsymbol{v}^{2,h}) \in V^h \mid v_\nu^{1,h} + v_\nu^{2,h} \leq 0 \text{ on } \Gamma_3\} \subset U.$$

Then, a fully discrete scheme, which is an improved version of the one in [12], is the following:

Problem 3.11 *Given* $\boldsymbol{u}_0^h \in U^h$ *and* $\boldsymbol{\sigma}_0^h \in Q^h$, *find a displacement field* $\boldsymbol{u}^{hk} = \{\boldsymbol{u}_n^{hk}\}_{n=0}^N \subset U^h$ *and a stress field* $\boldsymbol{\sigma}^{hk} = \{\boldsymbol{\sigma}_n^{hk}\}_{n=0}^N \subset Q^h$ *such that* $\boldsymbol{u}_0^{hk} = \boldsymbol{u}_0^h$, $\boldsymbol{\sigma}_0^{hk} = \boldsymbol{\sigma}_0^h$, *and, for* $n = 1, \ldots, N$,

$$\delta\boldsymbol{\sigma}_n^{hk} = \mathcal{P}_{Q^h}\mathcal{E}\delta\boldsymbol{\varepsilon}(\boldsymbol{u}_n^{hk}) + \mathcal{P}_{Q^h}G(\boldsymbol{\sigma}_n^{hk}, \boldsymbol{\varepsilon}(\boldsymbol{u}_n^{hk})),$$
$$(\boldsymbol{\sigma}_n^{hk}, \boldsymbol{\varepsilon}(\boldsymbol{v}^h - \boldsymbol{u}_n^{hk}))_Q \geq (\boldsymbol{f}_n, \boldsymbol{v}^h - \boldsymbol{u}_n^{hk})_V \quad \forall\, \boldsymbol{v}^h \in U^h.$$

By slightly modifying the arguments in [12], we have the following result.

Theorem 3.12 *Assume that the conditions in Theorem 3.10 hold and also the following ones:*

- $\|\boldsymbol{u}_0 - \boldsymbol{u}_0^h\|_V \leq ch, \quad \|\boldsymbol{\sigma}_0 - \boldsymbol{\sigma}_0^h\|_Q \leq ch$,
- $[H^2(\Omega^1)]^d \times [H^2(\Omega^2)]^d \cap U$ *is dense in* U *and*

$$\inf_{\boldsymbol{v}^h \in U^h} \|\boldsymbol{v} - \boldsymbol{v}^h\|_V \leq c(\boldsymbol{v})h, \quad \forall \boldsymbol{v} \in [H^2(\Omega^1)]^d \times [H^2(\Omega^2)]^d \cap U,$$

- $\|(I_Q - \mathcal{P}_{Q^h})\boldsymbol{\tau}\|_Q \leq ch, \quad \forall \boldsymbol{\tau} \in Q$,
- $\boldsymbol{u}^m \in L^\infty(0,T; [H^2(\Omega^m)]^d)\,(m = 1,2)$,

then

$$\max_{1 \leq n \leq N} \left(\|\boldsymbol{u}_n - \boldsymbol{u}_n^{hk}\|_V + \|\boldsymbol{\sigma}_n - \boldsymbol{\sigma}_n^{hk}\|_Q\right)$$
$$\leq ch^{1/2}\|\boldsymbol{u}\|_{L^\infty(0,T;[H^2(\Omega)]^d)} + c\,k\,\left(\|\dot{\boldsymbol{u}}\|_{L^\infty(0,T;V)} + \|\dot{\boldsymbol{\sigma}}\|_{L^\infty(0,T;Q)}\right).$$

If we further assume

- $u_\nu^m \in L^\infty(0,T; H^2(\Gamma_3))\,(m = 1,2), \quad \sigma_\nu \in L^\infty(0,T; L^2(\Gamma_3))$
- $\inf_{\boldsymbol{v}^h \in U^h} \left[\|\boldsymbol{v} - \boldsymbol{v}^h\|_V + \|v_\nu - v_\nu^h\|_{L^2(\Gamma_3)}^{1/2}\right] \leq c(\boldsymbol{v})h$, $\forall \boldsymbol{v} \in [H^2(\Omega^1)]^d \times [H^2(\Omega^2)]^d \cap U$,

then we have an optimal order error estimate

$$\max_{1 \leq n \leq N} \left(\|\boldsymbol{u}_n - \boldsymbol{u}_n^{hk}\|_V + \|\boldsymbol{\sigma}_n - \boldsymbol{\sigma}_n^{hk}\|_Q\right)$$
$$\leq ch\left(\|\boldsymbol{u}\|_{L^\infty(0,T;[H^2(\Omega)]^d)} + \|u_\nu\|_{L^\infty(0,T;H^2(\Gamma_3))}^{1/2}\right) + c\,k\,\left(\|\dot{\boldsymbol{u}}\|_{L^\infty(0,T;V)} + \|\dot{\boldsymbol{\sigma}}\|_{L^\infty(0,T;Q)}\right). \quad (27)$$

Note that these estimations are valid when $V^{m,h}$ is the space of piecewise polynomials of degree less or equal 1 and $Q^{m,h}$ is the space of piecewise constant functions.

For brevity, in this resumed version we do not include the analysis of interesting method of discretisation with non matching methods (see [8], [13]).

4 Numerical examples

To verify the accuracy of the numerical methods described in Section 3, a number of numerical experiments have been performed on test problems in one, two and three dimensions. We describe in this section some numerical results.

4.1 The Signorini contact problem in viscoplasticity

4.1.1 A one-dimensional test problem

Problem 3.1 has been tested with the data:

$$\Omega = (0,1), \quad T = 10 \text{ sec.}, \quad \Gamma_1 = \{0\}, \quad \Gamma_2 = \emptyset, \quad \Gamma_3 = \{1\}, \quad f_0(x) = 10N/m, \quad g = 0.25m,$$
$$u_0(x) = 0m, \quad \sigma_0(x) = 10 - 10x\, N/m, \quad \mathcal{E}(x) = 10N, \quad G(\sigma,\varepsilon) = -\sigma + 10\varepsilon.$$

In Section 3.1, we considered the Signorini contact problem with a zero gap. The results stated there can be extended straightforward to the situation with a nonzero initial gap g.

The exact solution of the 1-D problem is:

$$\text{For } 0 \le t \le \ln 2 \text{ (no contact)} \ : \ \begin{cases} \sigma(t,x) = 10 - 10x, \\ u(t,x) = (1-e^{-t})(x - \frac{x^2}{2}). \end{cases} \tag{28}$$

$$\text{For } t > \ln 2 \text{ (in contact)} : \begin{cases} \sigma(t,x) = \dfrac{5}{2}(2e^{-t}+3) - 10x, \\ u(t,x) = \dfrac{1}{2}x^2(e^{-t}-1) + \dfrac{1}{4}x[2e^{-t} + 3 - 4e^{-t}]. \end{cases} \tag{29}$$

Employing the fully discrete problem described in Section 3.1, the numerical method has been implemented. In Fig. 1, the displacement fields at the times $t = 0.5,\ 1,\ 2,\ 4,\ 8$ sec. are depicted. The discretization parameters are $k = 0.01$ and $h = 0.01$. The difference between the numerical and exact solutions (28)-(29) is also plotted.

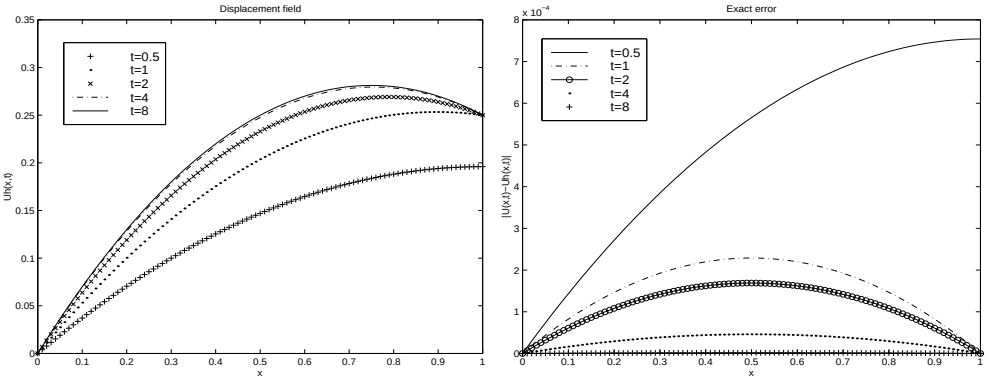


Figure 1: Displacement field and exact error for different time values.

In Figure 2 we show the evolution of the displacements at the nodes $x = 0.25,\ 0.5,\ 1$, and its corresponding error values. We observe the effect produced at the contact time $t = \ln 2$ (approx 0.69). Finally, the values of the exact error are calculated for a number of time and spatial discretization parameters, and asymptotic behaviour (16) has been obtained, for an asymptotic constant $C = 0.9645 \times 10^{-1}$.

4.1.2 A two-dimensional test problem

We use the data:

$$\Omega = (0,1)\times(0,1), \quad T = 1sec, \quad \Gamma_1 = [0,1]\times\{1\}, \quad \Gamma_2 = \{0,1\}\times(0,1), \quad \Gamma_3 = [0,1]\times\{0\},$$
$$\boldsymbol{f}_0 = (0,-10t)N/m^2, \quad \boldsymbol{f}_2 = (0,0)N/m, \quad \boldsymbol{\sigma}_0 = \mathbf{0}N/m^2, \quad \boldsymbol{u}_0 = \mathbf{0}m.$$

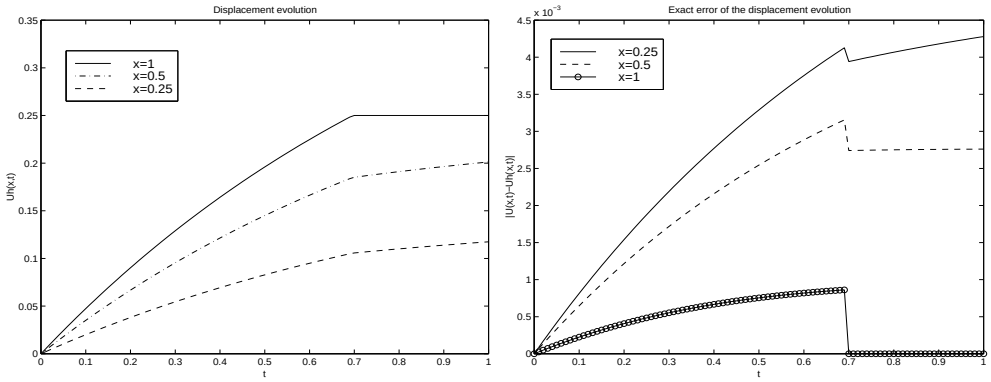


Figure 2: Evolution of displacements of points $x = 0.25, 0.5, 1$, and its corresponding exact error

$\mathcal{E}$ is the plane stress elasticity tensor:

$$(\mathcal{E}\boldsymbol{\tau})_{\alpha\beta} = \frac{E\kappa}{1-\kappa^2}(\tau_{11}+\tau_{22})\delta_{\alpha\beta} + \frac{E}{1+\kappa}\tau_{\alpha\beta},$$

for $\alpha, \beta = 1, 2$, where E is the Young's modulus and κ is the Poisson's ratio. In this example $E = 10^8 N/m^2$ and $\kappa = 0.3$.

We consider an obstacle defined implicitly by

$$\frac{(x_1-3)^2}{900} + \frac{(x_2+3)^2}{9} - 1 = 0,$$

and the gap function $g(\boldsymbol{x})$ is given as the distance between the contact point $\boldsymbol{x}$ and the obstacle.

The classical Perzyna's viscoplastic function (see [2] and [15]) has been considered in its 2-D version, i.e.,

$$G(\boldsymbol{\sigma}, \boldsymbol{\varepsilon}) = -\frac{1}{2\mu^*}\mathcal{E}(\boldsymbol{\sigma} - P_K\boldsymbol{\sigma}), \tag{30}$$

where $\mu^* > 0$ is the viscosity coefficient and P_K is the orthogonal projection operator (with respect to the norm $||\boldsymbol{\tau}|| = (\mathcal{E}\tau, \tau)^{1/2}$) over the convex subset $K \subset \mathbb{S}^2$ defined by:

$$K = \{\boldsymbol{\tau} \in \mathbb{S}^2 \mid \tau_{11}^2 + \tau_{22}^2 - \tau_{11}\tau_{22} + 3\tau_{12}^2 \leq \sigma_Y^2\},$$

σ_Y being the uniaxial yield stress. In this case, we used $\sigma_Y = \sqrt{10}N/m^2$ and $\mu^* = 100N/m^2$.

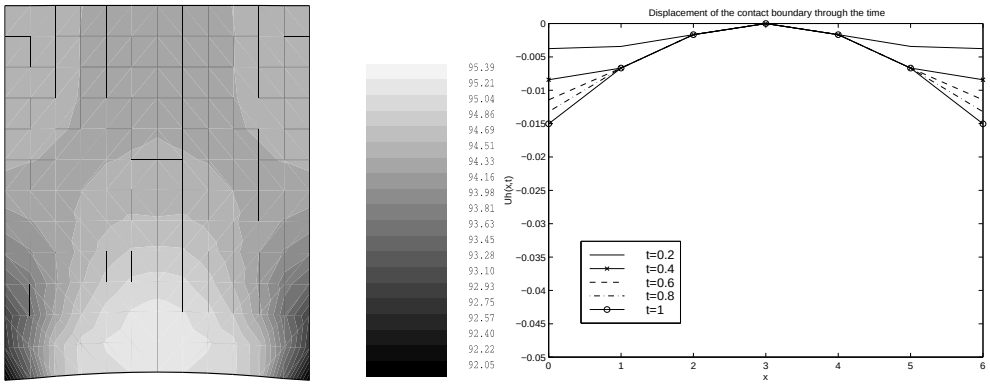


Figure 3: Von-Mises stress on deformed configuration and evolution of the u_2 component in a 2-D Signorini problem

In Figure 3 the deformed configuration, the Von-Mises norm for the stress at time $t = 1sec.$, and the evolution of the contact boundary are plotted.

4.1.3 A three-dimensional test problem

In the three-dimensional case, we use the data:

$$\Omega = (0,3)\times(0,1)\times(0,1), \quad T = 1\,sec, \quad \Gamma_1 = \{0\}\times[0,1]\times[0,1], \quad \Gamma_3 = (0,3)\times(0,1)\times\{0\},$$
$$\Gamma_2 = \Gamma - (\Gamma_1 \cup \Gamma_3), \quad \boldsymbol{\sigma}_0 = \mathbf{0}N/m^3, \quad \boldsymbol{u}_0 = \mathbf{0}m.$$
$$\boldsymbol{f}_2 = (0,0,0)N/m^2, \quad \boldsymbol{f}_0(x_1,x_2,x_3,t) = \begin{cases} (0,0,-10t)N/m^3 \text{ if } x_1 = 0, \\ (0,0,0)N/m^3 \text{ otherwise.} \end{cases}$$

Here, $\mathcal{E}$ is the three-dimensional elasticity tensor,

$$(\mathcal{E}\boldsymbol{\tau})_{ij} = \frac{E\kappa}{(1+\kappa)(1-2\kappa)}(\sum_{k=1}^{3}\tau_{kk})\delta_{ij} + \frac{E}{1+\kappa}\tau_{ij},$$

for $i,j = 1,2,3$, where Young's modulus E and Poisson's ratio κ are $10^8 N/m^3$ and 0.3, respectively. The constitutive function $G(\boldsymbol{\sigma},\boldsymbol{\varepsilon})$ is again Perzyna's, i.e. (30) in its three-dimensional version with

$$K = \{\boldsymbol{\tau} \in \mathbb{S}^3 \mid (\sigma_{11}-\sigma_{22})^2 + (\sigma_{22}-\sigma_{33})^2 + (\sigma_{33}-\sigma_{11})^2 + 6(\sigma_{12}^2+\sigma_{13}^2+\sigma_{23}^2) \le \sigma_Y^2\}.$$

Here, we used $\sigma_Y = \sqrt{10}N/m^3$. In Figure 4 the displacements and the Von-Mises norm for the stress are shown at the final time T. Also, the evolution of the second component of the displacement field of the contact nodes on surface $x_1 = 0$ is shown.

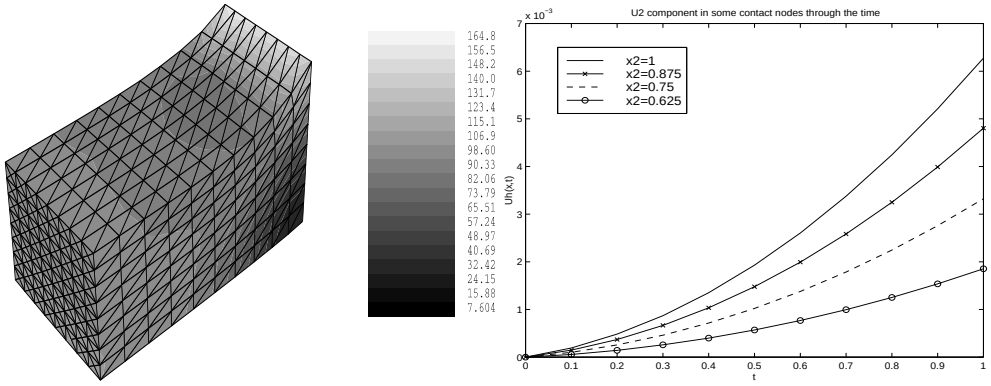


Figure 4: Von-Mises stress on deformed configuration and evolution of the u_2 component in a 3-D Signorini problem

4.2 Contact problem with normal compliance

Because of limited extension of the paper we only describe a one dimensional test. The contact problem with a deformable foundation described in Section 3.2 is considered with the data:

$$\Omega = (0,1), \quad T = 10\,sec, \quad \Gamma_1 = \{0\}, \quad \Gamma_2 = \emptyset, \quad \Gamma_3 = \{1\}, \quad f_0(x,t) = 10N/m, \quad g = 0.25m, \quad \alpha = 1,$$
$$u_0(x) = 0\,m, \quad \sigma_0(x) = 10 - 10x\,N/m, \quad \mathcal{E}(x) = 10N, \quad G(\sigma,\varepsilon) = -\sigma + 10\varepsilon, \quad r^* = 100N/m.$$

The exact solution of this problem is:

$$\text{For } 0 \le t \le \ln 2 \text{ (no contact)} : \begin{cases} \sigma(x,t) = 10 - 10x, \\ u(x,t) = (1-e^{-t})(x - \dfrac{x^2}{2}). \end{cases} \tag{31}$$

$$\text{For } t > \ln 2 \text{ (in contact)} : \begin{cases} \sigma(x,t) = \dfrac{5(2e^{-t}+3+40r^*)}{2(10r^*+1)} - 10x, \\ u(x,t) = \dfrac{x^2}{2}(e^{-t}-1) + x[\dfrac{2e^{-t}+3+40r^*}{4(10r^*+1)} - e^{-t}]. \end{cases} \tag{32}$$

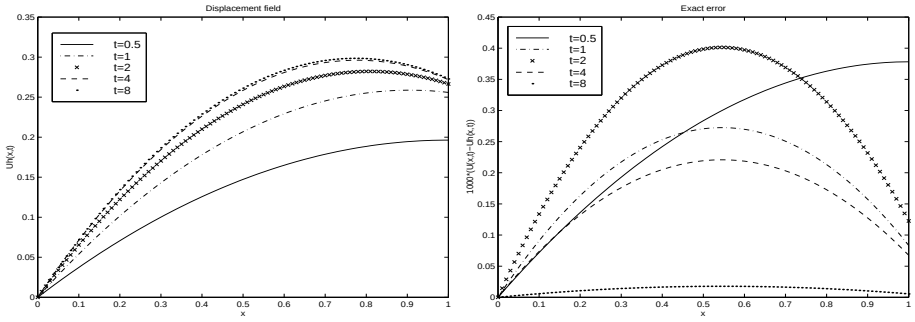


Figure 5: Displacement field and exact error at different times in a 1-D normal compliance problem.

By using the discrete problem in Section 5.2, we have implemented the numerical method on a standard workstation. Figure 5 depicts the displacements at the times $t = 0.5, 1, 2, 4, 8$ sec., calculated with parameters $h = 0.01$ and $k = 0.01$. We also plot the difference with the exact solution (32)–(32) scaled by the factor 10^3.

In Figure 6 we show the evolution of the points $x = 0.25,\ 0.5,\ 1$), and the corresponding error between the numerical solution and the exact values.

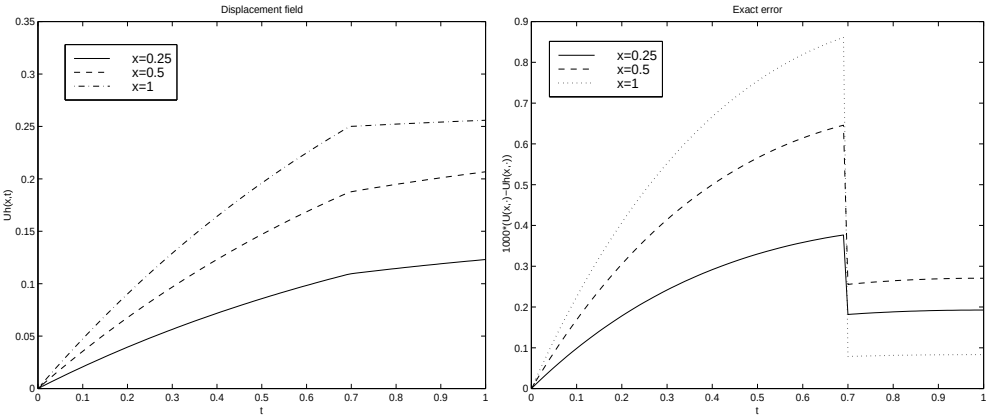


Figure 6: Evolution of displacements of points $x = 0.25, 0.5, 1$ and corresponding scaled exact error.

From the exact error values, asymptotic behaviour (21) is obtained with an asymtotic constant $C = 0.9557 \times 10^{-1}$, independent of h and k.

4.3 Contact problem between two viscoplastic bodies

4.3.1 A two-dimensional test problem: case 1

We consider the contact problem between two viscoplastic bodies described in Section 3.3 with the data:

$$\Omega^1 = (0,4)\times(0,1),\quad \Omega^2 = (0,4)\times(-1,0),\quad T = 1sec,\quad \Gamma_1^1 = \{4\}\times[0,1],\quad \Gamma_1^2 = \{4\}\times[-1,0],$$
$$\Gamma_3 = (0,4)\times\{0\},\quad \Gamma_2^1 = \Gamma^1 - (\Gamma_1^1\cup\Gamma_3),\quad \Gamma_2^2 = \Gamma^2 - (\Gamma_1^2\cup\Gamma_3),$$
$$\boldsymbol{f}_0^1 = (0,0)N/m^2,\quad \boldsymbol{f}_0^2 = (0,0)N/m^2,\quad \boldsymbol{f}_2^2 = (0,0)N/m,\quad \boldsymbol{\sigma}_0 = \mathbf{0}N/m^2,\quad \boldsymbol{u}_0 = \mathbf{0}\,m,$$

$$\boldsymbol{f}_2^1(x_1,x_2,t) = \begin{cases} (0,-10t)N/m \text{ if } 3\le x_1\le 4,\ x_2 = 1,\\ (10t,0)N/m \text{ if } 0.5\le x_2\le 1,\ x_1 = 0,\\ 0 \text{ otherwise,}\end{cases}$$

The Young's modulus and Poisson's ratio for the two viscoplastic bodies Ω^1 and Ω^2 are $10^8 N/m^2$ and $\kappa = 0.3$. The Perzyna law (30) is used, where $\mu^* = 100N/m^2$ and $\sigma_Y = 10N/m^2$. Figure 7 depicts the displacements and

the Von-Mises norm for stress at the final time.

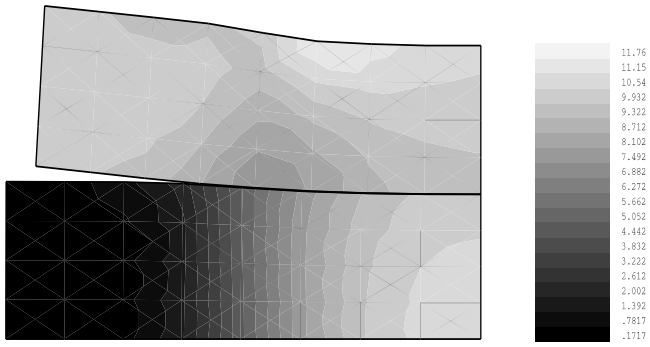


Figure 7: The displacements and the Von-Mises norm in a 2-D contact problem between two viscoplastic bodies (case 1).

4.3.2 A two-dimensional test problem: case 2

In this case, we study the contact problem between two viscoplastic bodies in the setting described in Figure 8 is considered.

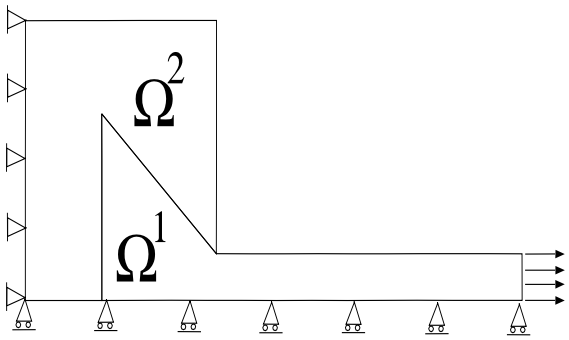


Figure 8: Contact between two viscoplastic bodies (case 2).

The Young's modulus and Poisson's ratio for the two viscoplastic bodies Ω^1 and Ω^2 are $10^8 N/m^2$ and $\kappa = 0.3$. The Perzyna law (30) is used, where $\mu^* = 100N/m^2$ and $\sigma_Y = 10N/m^2$. In Figure 9 the displacements and the Von-Mises norm for stress at the final time are shown.

4.3.3 A three-dimensional test problem

Finally, we consider a contact problem between two viscoplastic bodies in three dimensions. The following data have been used:

$$\Omega^1 = (1,2)\times(1,4)\times(0,1), \quad \Omega^2 = (0,3)\times(0,1)\times(0,1), \quad T = 1sec, \quad \Gamma_1^1 = \emptyset,$$
$$\Gamma_1^2 = \{0,3\}\times[0,1]\times[0,1], \quad \Gamma_3 = (1,2)\times\{1\}\times(0,1), \quad \Gamma_2^1 = \Gamma^1 - (\Gamma_1^1 \cup \Gamma_3), \quad \Gamma_2^2 = \Gamma^2 - (\Gamma_1^2 \cup \Gamma_3),$$
$$\boldsymbol{f}_0^1 = \begin{cases} (0,-100t,0)N/m^2 \text{ if } x_2 = 4,\ 1 \le x_1 \le x_2,\ 0 \le x_3 \le 1, \\ (0,0,0) \text{ in another case.} \end{cases}$$
$$\boldsymbol{f}_0^2 = (0,0,0)N/m^2, \quad \boldsymbol{f}_2^1 = (0,0,0), \quad \boldsymbol{f}_2^2 = (0,0,0)N/m, \quad \boldsymbol{\sigma}_0 = \mathbf{0}N/m^2, \quad \boldsymbol{u}_0 = \mathbf{0}m.$$

As above, Perzyna's law and elasticity tensor $\mathcal{E}$ were considered with parameters $\mu^* = 100N/m^3$, $E = 10^8\ N/m^2$ and $\kappa = 0.3$. Figure 10 depicts the displacemeyts and the Von-Mises norm for stress are shown.

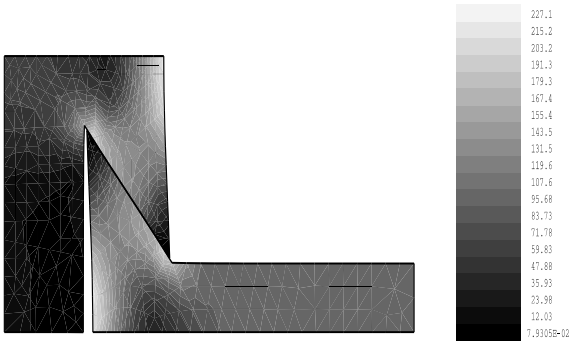


Figure 9: Displacements and the Von-Mises norm in a 2-D contact problem between two viscoplastic bodies (case 2).

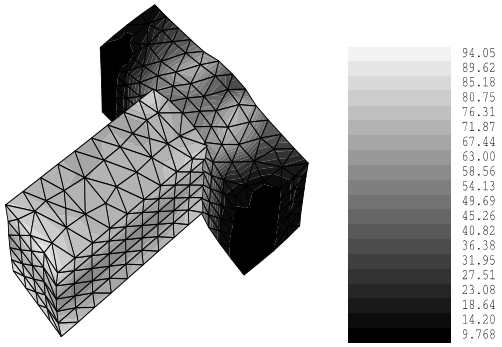


Figure 10: Displacements and the Von-Mises norm in a 3-D contact problem between two viscoplastic bodies.

5 Conclusion

Considerable progress has been made on the modelling, variational analysis and numerical analysis of quasistatic contact phenomena. Our understanding of the behavior of the models for these processes has deepened, and the new problems led to the investigation of new types of variational inequalities. These kind of methods can be extended to quasistatic and dynamic contact problems for viscoelastic materials (see, for example, [1], [20], [21]) and other models including damage and wear that leads to new and interesting types of variational inequalities (see [1], [10], [22], [25]).

References

[1] Campo M., Fernández J.R., Shillor, M. and Viaño J.M.: *Analysis and simulations of a dynamic viscoelastic contact problem with damage*. (Submitted).

[2] Cristescu N and Suliciu I.: Viscoplasticity. *Martinus Nijhoff Publishers*, Editura Tehnica, Bucharest, 1982.

[3] Chen J., Han W. and Sofonea M.: *Numerical analysis of a nonlinear evolution system with applications in viscoplasticity*. SIAM J. Numer. Anal. Vol. 38, pp. 1171-1199, 2000.

[4] Chen J., Han W. and Sofonea M.: *Numerical analysis of a contact problem in rate-type viscoplasticity*. Numer. Funct. Anal. Optim. Vol. 22, pp. 505-527, 2001.

[5] Ciarlet P.G.: The Finite Element Method for Elliptic Problems. *North Holland*, Amsterdam, 1978.

[6] Duvaut G. and Lions J.L.: Inequalities in Mechanics and Physics. *Springer-Verlag*, Berlin, 1976.

[7] Fernández-García J. R.: Análisis numérico de problemas de contacto sin rozamiento en viscoplasticidad. *Doctoral dissertation*. University of Santiago de Compostela, 2002.

[8] Fernández-García J. R., Hild P. and Viaño J.M.: *Numerical approximation of the elastic-viscoplastic contact problem with non-matching meshes* . Numer. Math. (To appear).

[9] Fernández-García J.R. , Sofonea M. and Viaño J. M.: *A frictionless contact problem for elastic-viscoplastic materials with normal compliance: numerical analysis and computational experiments*. Numer. Math. Vol. 90, pp. 689-719, 2002.

[10] Fernández-García J.R., Sofonea M. and Viaño J. M.: *Numerical analysis of a quasistatic viscoelastic sliding frictional contact problem with wear*. Comput. Methods in Appl. Mech. Engrg. (To appear).

[11] Fichera G.: *Problemi elastostatici con vincoli unilaterali. II. Problema di Signorini con ambique condizioni al contorno*, Mem. Accad. Naz. Lincei, S. VIII, Vol. VII, Sez. I, 5, pp. 91–140, 1964.

[12] Han W. and Sofonea M.: *Numerical analysis of a frictionless contact problem for elastic-viscoplastic materials*. Comput. Methods Appl. Mech. Engrg. Vol 190, pp. 179-191, 2000.

[13] Hild P.: *Numerical implementation of two nonconforming finite element methods for unilateral contact*. Comput. Methods Appl. Mech. Engrg. Vol 184, pp. 99-123, 2000.

[14] Hlaváček I., Haslinger J., Necăs J. and Lovíšek J.: *Solution of Variational Inequalities in Mechanics*. Springer-Verlag, New York, 1988.

[15] Ionescu I.R. and Sofonea M.: Functional and Numerical Methods in Viscoplasticity. *Oxford University Press*, Oxford, 1993.

[16] Kikuchi N. and Oden J.T.: Contact Problems in Elasticity: A Study of Variational Inequalities and Finite Element Methods. *SIAM*, Philadelphia, 1988.

[17] Raous M., Jean M. and Moreau J.J. (eds.): Contact Mechanics. *Plenum Press*, New York, 1995.

[18] Rochdi M. and Sofonea M.: *On frictionless contact between two elastic-viscoplastic bodies*. Quart. J. Mech. Appl. Math., Vol. 50, pp. 481-496, 1997.

[19] Rochdi M., Sofonea M. and Rosca I.: *On existence, behaviour, and numerical approach of the solution for contact problems in rate-type viscoplasticity*. In Proceedings of the 1996 Engineering Systems Design and Analysis Conference, ESDA'96, eds. A.B. Sabir et al. pp. 87–92, 1996.

[20] Rodríguez-Arós A.D., Sofonea M. and Viaño J.M.: *A class of evolutionary variational inequalities with Volterra-type term*. Math. Mod. Meth. Appl. S. (M3AS) (to appear).

[21] Rodríguez-Arós A.D. , Sofonea M. and Viaño J.M.: *A class of integro-differential variational inequalities with applications in frictional contact problems in viscoelasticity*. (Submitted).

[22] Shillor M. (ed.): *Recent Advances in Contact Mechanics*, Math. Computer Model., Vol. 28 (4–8), 1998.

[23] Signorini A.: *Sopra a une questioni di elastostatica.* Atti della Società Italiana per il Progresso delle Scienze, 1933.

[24] Sofonea M.: *On a contact problem for elastic-viscoplastic bodies*. Nonlinear Analysis, Theory, Methods and Applications, Vol. 29, pp. 1037–1050, 1997.

[25] Sofonea M. and Shillor M.: *Variational analysis of quasistatic viscoplastic contact problems with friction*. Commun. Appl. Anal. Vol. 5, pp. 135-151, 2001.

Numerical analysis of an integral equation modeling a radiative transfer problem

Filomena Dias d'Almeida
CMUP-Centro de Matemática da Universidade do Porto e
Faculdade de Engenharia da Universidade do Porto,
rua Roberto Frias s/n, 4200-465 Porto, Portugal.
(falmeida@fe.up.pt)

The example to be shown concerns a radiative transfer equation that occurs in Astrophysics. We will consider a restricted problem obtained when the temperature and pressure are known in which case the system becomes linear. The transfer problem of the absortion of photons in stellar atmospheres may be described by a Fredholm integral equation of the second kind, weakly singular. The problem will be set in the Banach space $L^1(I)$.

The numerical approximation is based on a sequence of projections onto successive finite dimensional subspaces spanned by n functions piecewise constant in each of the subintervals determined by a non uniform grid of $n+1$ points in I.

To obtain a precision equivalent to the application of the previous projection method on a large dimensional subspace without the solution of the corresponding large linear system, we will use iterative refinement formulae. These are to be applied to an approximate solution obtained with a system of small dimension.

Sparse matrix techniques and parallel processing are also used to accelerate the computations.

(This is a joint work with P. B. Vasconcelos, M. Ahues, A. Largillier, O. Titaud, B. Rutily)

A superconvergent piecewise linear finite element method for elliptic system of partial differential equations

S.Barbeiro
Universidade de Coimbra, Faculdade de Ciências e Tecnologia,
Departamento de Matemática, Apartado 3008, 3000 Coimbra, Portugal.
email: silvia@mat.uc.pt
J.A. Ferreira
Universidade de Coimbra, Faculdade de Ciências e Tecnologia
Departamento de Matemática, Apartado 3008, 3000 Coimbra, Portugal.
email: ferreira@mat.uc.pt

Abstract

Most quantities appearing in physical applications are ruled by systems of partial differential equations. An example is given by the deformations and stresses of elastic and inelastic bodies subject to load, studied in solid mechanics.

It is well known that for problems defined in two dimensional domains, piecewise linear finite element solutions are second order approximations for the solution with respect to $\mathbf{L_2}$ norm, but their gradient are only first order approximations for the gradient of the solution. These convergences are obtained assuming that the triangulations of the domain are quasi-uniform and regular.

In this talk we study the convergence properties of the numerical approximations for the solution of systems of elliptic equations defined on two dimensional polygonal domains. These approximations are constructed using a non standard fully discrete piecewise linear finite element method based on non uniform triangulations and considering a variational formulation with a sesquilinear form which can be not strongly coercive. For $s \in \{1, 2\}$, we prove order s convergence for the piecewise linear finite element solution and its gradient, if the solution of the system is in the Sobolev space $\mathbf{H}^{s+1}(\Omega)$.

Several authors studied the superconvergence of the gradient. For instance, about two decades ago, M. Zlámal found superconvergence of the gradient for certain quadrature finite element solutions on nearly rectangular grids. Furthermore J. Brandts studied superconvergence of the gradient of the piecewise linear finite element solution, but the grids were assumed regular and quasi-uniform.

The nonstandard finite element method studied in this work is equivalent to a carefully defined finite difference method and hence we conclude that this last method is supraconvergent. Supraconvergent finite difference schemes have been largely studied in the literature.

In the present work we start describing the nonstandard piecewise linear finite element method for a general uniformly strongly elliptic system. The stability of the sesquilinear form that defines the nonstandard method is established. Using the stability properties we study the behavior of the error. Examples illustrating the performance of the method are considered for planar elasticity problems.

Advanced School and Workshop on Modelling and Numerical Simulation in Continuum Mechanics
14-18 July, 2003, Coimbra, Portugal

On the motion of a particle in vortical flows

Marcelo H. Kobayashi

Instituto Superior Técnico, Department of Mechanical Engineering/SMA
Av. Rovisco Pais, 1, 1049–001 Lisboa, Portugal
email: marcelo@popsrv.ist.utl.pt

In this lecture, the dynamics of a non-neutrally buoyant particle moving in a rotating vessel filled with a Newtonian fluid is examined analytically and experimentally. The geometry under study is used in mineral particle separators, suspension bioreactors and other equipment of industrial interest. In particular, the dynamical problem to be discussed simulates the motion of micro-carriers in NASAs microgravity bioreactors. These rotating vessel bioreactors have been used extensively to study suspended tissue growth on earth. Virtual mass, gravity, pressure, steady and history drag effects at low particle Reynolds numbers are considered. The presence of lift forces, both steady and unsteady, are taken into account. Results are compared to traditional formulations of low Reynolds flows that do not account for small, inertial lift effects. Substantial differences were found by including lift in the formulation during our preliminary analysis and therefore we seek to confirm these exciting results experimentally. For particles lighter than the fluid, an asymptotically stable equilibrium position was found to be at a horizontal distance from the center of rotation and at an angle with the X-axis.

To our knowledge this work is the first to solve the particle Lagrangian equation of motion in its complete form (with or without lift) for a non-uniform flow using an exact method, and also the first to validate relevant expressions for Saffmans and McLaughlins lift coefficients with this flow configuration. Our formulation of this problem predicts that even at very small rotation rates a remarkable phenomenon occurs due to lift effects exclusively: the equilibrium position of particles lighter than the fluid is always below (assuming the gravity acceleration to point down) the horizontal plane containing the axis of the cylinder. This result is in direct contrast with the behavior predicted by the Maxey-Riley equation which does not include lift effects. The exact solution of the Maxey-Riley equation derived during this research effort predicts that a light particle will reach equilibrium above the central plane. We will show that even at very small Rep and shear Reynolds number Res lift effects will force a light particle below the central horizontal plane, provided that Rep is less or equal to Res. We also show that this flow configuration can be used to determine experimentally the lift coefficient for a particle in a uniform vorticity field.

APPLICATION OF NUMERICAL AND EXPERIMENTAL TECHNIQUES IN BULK FORMING

P. A. F. Martins
Departamento de Engenharia Mecânica, Instituto Superior Técnico, Lisboa
pmartins@ist.utl.pt

ABSTRACT

Bulk forming is a critical activity characterized by short lead times and constant technological modifications in order to improve quality and reduce manufacturing costs.

The utilization of experimental techniques and numerical simulation softwares at both basic and advanced levels can help engineers solving different technological tasks; (i) they may be used as tools for designing and optimizing a process, (ii) they may help testing the impact of different raw materials and lubricants on the final properties of the formed parts, and (iii) they may also serve in-plant engineers debugging and solving formability problems, evaluating possible changes in process parameters and making small modifications in the shape of already existing dies/tools.

This presentation outlines a number of examples and discusses the benefits and limitations in using experimental and numerical simulation procedures.

EVOLUTIONARY ALGORITHMS IN MULTIOBJECTIVE OPTIMIZATION

Pedro Oliveira

Departamento de Produção e Sistemas
Escola de Engenharia
Universidade do Minho
4800 Guimarães
pno@dps.uminho.pt

Solving multiobjective engineering problems is a very difficult task due to, in general, in this class of problems, the objectives conflict across a high-dimensional problem space. In these problems, there is no single optimal solution; the interaction of multiple objectives gives rise to a set of efficient solutions, known as the Pareto-optimal solutions. During the past decade, Genetic Algorithms (GAs) (Goldberg, 1989) were extended in order to tackle this class of problems, such as the work of Schaffer (1985), Fonseca and Fleming (1995), Horn et al. (1994), Srinivas and Deb (1995) and, Zitzler and Thiele (1998). These multiobjective approaches explore some features of Evolutionary Algorithms, in particular, since these algorithms work with populations of candidate solutions, they can, in principle, find multiple Pareto-optimal solutions in a single run; on the other hand, using some diversity-preserving mechanisms Evolutionary Algorithms can find widely different Pareto-optimal solutions. In this talk a review of the latest developments on evolutionary multiobjective optimization is going to be presented, with some examples on structural optimization.

SOME IMPROVEMENTS IN MODELLING OF METAL FORMING PROCESSES INVOLVING ELEMENT ARCHITECTURE, CONTACT DETECTION AND DAMAGE

José M. A. César de Sá & P.M.A. Areias
IDMEC- Faculdade de Engenharia da Universidade do Porto
Rua Dr. Roberto Frias s/n Porto, Portugal
cesarsa@fe.up.pt; http://www.fe.up.pt

ABSTRACT

Metal forming processes are generally characterised for involving important changes of the initial shape of a workpiece by plastic deformation controlled by contact, with friction, provided by the tools. These processes have been generating challenging problems in the numerical modelling. Some developments in the solution of some of those problems are here addressed.

In order to handle the large deformations involved in metal forming and the "locking" problems related to the incompressibility of plastic deformation a lot of effort has been put forward on element architecture. One of the more recent approaches is based on the concept of enhanced strain element that was established as a generalisation of the so-called incompatible modes element. The classical enhanced strain elements presented severe hourglass instabilities in certain finite strain regimes. Departing from the classical enhanced strain technique a special 3-D hexahedral element is described. The formulation contains a penalty stabilizing term that results naturally from the variational principle associated to the total potential as defined in the original formulation but not assuming orthogonallity between enhanced strains and stresses. The element performs well in large finite strain problems, has no special treatment or directional enhancement in order to be used in plate and shell analysis and may be used in bulk forming, solid analysis, sheet metal forming and in classical beam and plate analysis.

The contact problem between deformable bodies in 3-D, including friction, is still a challenging problem to be solve adequately in large-scale applications. In metal forming processes this may be an important issue to be addressed if the tool deformation or tool ware are to be taken into account or if self-contact in workpiece regions is to be avoided. Some innovative procedures are presented for contact detection and the circumventing of the equidistance dilemma and face selection, including criteria for avoiding wrong selection of target faces. A new Augmented Lagrangian function corresponding to a variation of the classical Rockafellar Lagrangian is proposed resulting in continuous second order derivatives if Lagrange multipliers are greater or equal than one and therefore avoiding sequential unconstrained minimization techniques. A new regularisation approach for friction forces, which rely solely on the use of curvilinear coordinates rather than a particular stress rate, is proposed.

An important aspect in the analysis of finite strain plasticity and metal forming is the one related to the representation of material softening behaviour and particularly strain localisation, either shear bending or localised necking that are known to pre-date ductile failure. A continuous damage model in close coupling with finite strain plasticity may numerically represent ductile fracture mechanisms. The straightforward numerical implementation of the softening part of ductile material behaviour, which may theoretically represent discontinuous strain rate components and shear bands leads to mesh size and orientation dependence. A gradient damage model is proposed with the purpose of attenuating mesh dependency. The stain softening behaviour is modelled through a variant of Lemaitre's damage evolution law, which allows taking into account the crack closure effect in compression.

Some Contact Problems in Viscoelasticity with Long-term Memory

Ángel Rodríguez-Arós
Departamento de Matemática Aplicada, Univ. de Santiago de Compostela, Spain
M. Sofonea
Laboratoire de Théorie des Systèmes, Université de Perpignan, France
J.M. Viaño
Departamento de Matemática Aplicada, Univ. de Santiago de Compostela, Spain
angelaros@usc.es sofonea@univ-perp.fr maviano@usc.es

Contact phenomena involving deformable bodies abound in industry and everyday life. The contact of the braking pads with the wheel or the tire with the road are two simple examples. In the last decades, a considerable progress has been made in their modelling and analysis, and the literature in this field is extensive. See, for example, [2] for a survey devoted to the contact of elastic bodies.

We investigated recently two frictionless models, for the Signorini contact and for the contact with normal compliance (see [3, 4]). In these papers the material was assumed to have linear viscoelastic behavior with long-term memory, that we describe with a Volterra-type integral equation of the form

$$\sigma_{ij}(t) = \mathcal{A}_{ijkl}\varepsilon_{kl}(\boldsymbol{u}(t)) + \int_0^t \mathcal{B}_{ijkl}(t-s)\varepsilon_{kl}(\boldsymbol{u}(s))ds,$$

where $\boldsymbol{\sigma} = (\sigma_{ij}), \boldsymbol{u} = (u_i)$ and $\boldsymbol{\varepsilon}(\boldsymbol{u}) = (\varepsilon_{ij}(\boldsymbol{u}))$ represent the stress tensor, the displacement field and the linearized strain tensor, respectively. Moreover, $\mathcal{A} = (\mathcal{A}_{ijkl})$ and $\mathcal{B} = (\mathcal{B}_{ijkl})$ are the fourth order tensors of elastic coefficients and the relaxation tensor, respectively. Thus, at each time $t > 0$ the stress tensor depends on all the previous strain states. Real materials in nature, like rubbers, organic polymeers or some kinds of wood have such a mechanical behavior.

In a variational form, the mechanical problems studied in the above papers lead to evolutionary variational inequalities for the displacement field involving an integral term of Volterra type. Here, we will show that, as the obstacle becomes less deformable, the weak solution of the normal compliance contact problem tends to the weak solution of the Signorini contact problem. We also present some numerical results of simulations that confirm the theoretical exposition.

Finally, we improve our model by taking into account the Tresca's friction law. This leads to an evolutionary variational inequality involving both a Volterra-type integral term and a partial derivative term with respect to the time variable. The analysis has been performed by using arguments of evolutionary inequalities established in [1], convexity and fixed point.

References

[1] H. Brézis, *Problèmes unilatéraux*, J. Math. Pures et Appl. **51** (1972), 1–168.

[2] N. Kikuchi and J. T. Oden, *Contact Problems in Elasticity: A Study of Variational Inequalities and Finite Element Methods*, SIAM, Philadelphia, 1988.

[3] A. Rodríguez-Arós, M. Sofonea and J. M. Viaño, A Signorini Frictionless Contact Problem for Viscoelastic Materials with Long-term Memory, in F. Brezzi, A. Buffa (Eds.) *Proceedings of European Conference on Numerical Mathematics and Advanced Applications*, Springer-Verlag, to appear.

[4] A. Rodríguez-Arós, M. Sofonea and J. M. Viaño, A Class of Evolutionary Variational Inequalities with Volterra-type Term, *Math. Mod. Meth. Appl. S.*, to appear.

Analysis and optimization of mixtures of materials at the microscale level

Cristian Barbarosie

Centro de Matemática e Aplicações Fundamentais
Av. Prof. Gama Pinto, 2, 1649-003 Lisboa
e-mail barbaros@ptmat.fc.ul.pt

A heterogeneous medium is considered, occupying the entire plane $\mathbb{R}^2$, with periodic heterogeneities. The heat conduction problem and the linear elasticity problem are studied in this infinite medium. This is done numerically by implementing a periodic finite element mesh, which can also be viewed as a mesh on the two-dimensional torus.

The goal of this study is to compute the effective conductivity/elasticity tensor, which describes the overall properties of the body when charges are applied "at infinity". Then, shape optimization is performed: one looks for the geometry of the heterogeneities that optimizes, in some sense, the effective properties of the body. A matrix made of a certain material is considered, having periodically distributed inclusions of a weaker material. The analysis is restricted to one periodicity cell only; one or two inclusions are considered in the cell. The shape of these inclusions is changed gradually in order to optimize the effective properties of the body.

A functional depending on the effective coefficients and on the volume proportion is defined, and a minimization algorithm is applied to this functional. One needs to compute the derivative of the functional with respect to the shape of the inclusion(s). This information is used by the minimization algorithm (a steepest descent algorithm).

The finite element mesh must change its geometry and topology along the optimization process. The inclusions can pass through the border of the periodicity cell, as long as they do not touch the other inclusions, or their own translations.

A viscoelastic beam oscillating between two stops with damage

M. Campo[1] J.R. Fernández[1] and M. Shillor[2]

In many materials there is an important decrease in their load bearing capacity, because of development of internal microcracks. As a result of the tensile or compressive stresses in the body, these microcracks open and grow which, in turn, causes the load bearing of the material to decrease. This reduction in the strenght of the material is modelled by introducing the damage field $\beta = \beta(x,t)$ as the ratio

$$\beta = \beta(x,t) = \frac{E_{eff}}{E}$$

between the effective modulus of elasticity E_{eff} and the modulus of the damage-free material E. Following Frémond and Nedjar [**?**], the evolution of the microscopic cracks causing the damage is described by the differential inclusion ([**?**, **?**])

$$c_d\beta' - \kappa\beta_{xx} - m\left(\frac{1-\beta}{\beta}\right) + d_1(u_{xx})_+^2 + d_2(u_{xx})_-^2 - q \in \partial\chi_{[\beta_*,1]}(\beta),$$

where $\kappa > 0$ is a constant relating to the diffusion of damage and c_d, m, d_1, d_2, q and β_*, $0 < \beta_* < 1$, are process parameters that must be obtained experimentally.

In the present work, we consider an uniform viscoelastic beam which is clamped at one of its ends to an oscilating device. The motion of the other end is constrained by two obstacles: the stops. This problem, without considering the damage, was introduced in [**?**]. The contact was supposed to be without friction and was modelled with a normal compliance condition, i.e., the stops are assumed to be flexible, with resistance proportional to the deflection.

A fully discrete scheme is proposed for the numerical solution of the model, using the finite element method to approximate the spatial variable and the Euler method to discretize the time derivatives. Error estimates are derived for the approximative solutions. The scheme was implemented on computer and numerical simulations of the evolution of the mechanical state and damage of the material will be presented.

References

[1] K. T. Andrews, K. Kuttler, M. Rochdi and M. Shillor, One-dimensional dynamic thermoviscoelastic contact with damage, *J. Math. Anal. Appl.* **272**, (2002), 249-275.

[2] M. Frémond and B. Nedjar, Damage, gradient of damage and principle of virtual power, *Internat. J. Solids Structures* **33** (8)(1996), 1083–1103.

[3] K. L. Kuttler and M. Shillor, Vibrations of a beam between two stops, *Dyn. Contin. Discrete Impuls. Syst. Ser. B Appl. Algorithms* **8** (2001), 93–110.

1 Departamento de Matemática Aplicada. Universidade de Santiago de Compostela. Facultade de Matemáticas s/n, Campus Sur. 15782 Santiago de Compostela, España.

2 Department of Mathematics and Statistics. Oakland University. Rochester, MI 48309, EEUU.

Instability and Bifurcation Modes in Systems with Coulomb Friction

A. Pinto da Costa and J.A.C. Martins
Instituto Superior Tcnico, Dep. Eng. Civil e Arquitectura and ICIST
Av. Rovisco Pais, 1049-001 Lisboa, Portugal
e-mails: apcosta,jmartins@civil.ist.utl.pt, Tels.: (351) 21 8418408(11)

The study of instabilities and bifurcations in systems with friction has been motivated by several experimental observations related to technological problems and industrial processes (like the occurrence of break squeal in vehicles or intermittent flows in granular media). The scientific community has also become progressively aware that friction-induced instabilities are responsible for the occurrence of earthquakes.

The research summarized in this presentation addresses two phenomena that, under certain conditions, may occur in mechanical systems with unilateral contacts with friction: (i) the directional instability of static equilibrium configurations; this is a dynamic (divergence) instability phenomenon caused by combined *stiffness*, *mass* and *friction* effects; (ii) the occurrence of angular bifurcations in quasi-static trajectories; this is a case of multiplicity of quasi-static solutions caused by combined *stiffness* and *friction* effects. We deal with finite dimensional plane linearly elastic systems constrained by plane rigid frictional contacts.

The stability study leads to a complementarity eigenproblem and we use it to give a necessary and sufficient condition for an equilibrium state to be directionally unstable. The instability modes and the corresponding coefficients of friction at the stability-instability transition may be obtained by solving another complementarity eigenproblem in which the coefficient of friction is the unknown eigenvalue.

Another kind of problem that may be formulated at an equilibrium state is the quasi-static rate problem, which consists of finding the first order right rates of change of displacements and reactions, for a given external loading rate at that state. For plane systems this problem may be formulated as a linear complementarity eigenproblem.

For constant external forces at a given equilibrium state, the directional instability problem at the stability-instability transition and the rate problem are the same. This means that an eigenmode in the stability-instability transition corresponds to an infinity of solutions to the rate problem for constant applied loads, and vice-versa.

For several finite element discretizations of elastic solids in frictional contact with flat obstacles, we show, compare and discuss the solutions to the above problems that are computed with two different algorithms.

Link between the Kirchhoff-Love and the Bernoulli-Navier models for a rectangular plate/beam

Carolina Ribeiro
Departamento de Matemática para a Ciência e a Tecnologia, Universidade do Minho,
Campus de Azurém, 4800-058 Guimarães, Portugal.
E-mail: cribeiro@mct.uminho.pt

Juan M. Viaño
Departamento de Matemática Aplicada, Universidade de Santiago de Compostela,
Faculdade de Matemáticas, 15782 Santiago de Compostela, Spain.
E-mail: maviano@usc.es

Abstract

Let $\Omega^{\varepsilon t} = (0, L) \times (-t, t) \times (-\varepsilon, \varepsilon)$ be the reference configuration of an elastic, homogeneous and isotropic solid. We assume ε and t to be very small with respect to L (length), so that $\Omega^{\varepsilon t}$ can be seen as a plate of thickness 2ε and middle surface $(0, L) \times (-t, t)$ or as beam with cross section $(-t, t) \times (-\varepsilon, \varepsilon)$ (which has area $A^\tau = 4\varepsilon t$). The body $\Omega^{\varepsilon t}$ is assumed to be clamped in one or two ends $\{0, L\} \times [-t, t] \times [-\varepsilon, \varepsilon]$. The plate/beam is submitted to the action of volume forces and surface tractions acting only on the upper and lower faces $[0, L] \times [-t, t] \times \{-\varepsilon, \varepsilon\}$. The part of the lateral surface not clamped is assumed free of forces.

We denote by $\boldsymbol{u}^{\varepsilon t}$ the corresponding displacement field, solution of the three-dimensional linear elasticity model. The plates theory justifies that for ε sufficiently small $\boldsymbol{u}^{\varepsilon t}$ can be approximated by $\bar{\boldsymbol{u}}^{\varepsilon t}$ where the bending $\bar{u}_3^{\varepsilon t}$ is the solution of the Kirchhoff-Love model and $(\bar{u}_1^{\varepsilon t}, \bar{u}_2^{\varepsilon t})$ solves a plane elasticity problem, both problems posed in the middle surface. A mathematical justification of this fact is now well-known [1]. In the same way, the beams theory for this case proposes to approximate $\boldsymbol{u}^{\varepsilon t}$ by $\tilde{\boldsymbol{u}}^{\varepsilon t}$, where the flexions $(\tilde{u}_2^{\varepsilon t}, \tilde{u}_3^{\varepsilon t})$ are the solution of the model of Bernoulli-Navier and $\tilde{u}_1^{\varepsilon t}$ is determined from the stretching equation, both problems posed on the interval $(0, L)$. This approach is also mathematically justified by asymptotic analysis [3].

In this work we assume that ε and t are of the same order of magnitude and we try to answer the following question: what is the rapport between $\bar{\boldsymbol{u}}^{\varepsilon t}$ (Kirchhoff-Love solution) and $\tilde{\boldsymbol{u}}^{\varepsilon t}$ (Bernoulli-Navier solution). Given the linearity of the equations of the Kirchhoff-Love model, we impose conditions to the material and choose appropriate assumptions the order magnitude of the forces. Using the asymptotic technique (Lions [2]) taking t as small parameter on the Kirchhoff-Love model, after a change of variable (consisting of a zoom in x_2-direction) to the reference middle surface $(-1, 1) \times (-\varepsilon, \varepsilon)$ and a suitable scaling of the unknowns, we prove that, up a factor $1 - \nu^2$ (ν: Poisson's coefficient of the material), the (scaled) Bernoulli-Navier model is the $H^1 \times H^1 \times H^2$-limit of the (scaled) Kircchoff-Love model as t tends to zero. In other words, de Bernoulli-Navier model is the natural approximation of the Kirchhoff-Love model when t is sufficiently small.

References

[1] P.G. Ciarlet, *Mathematical Elasticity*, Vol. II: *Theory of Plates*, North-Holland (1997).

[2] J. L. Lions, *Perturbations Singulières dans les Problèmes aus Limites et en Contrôle Optimal*, Lectures Notes in Mathematiques, Vol. 323, Springer-Verlag (1973).

[3] L. Trabucho and J. M. Viaño, Mathematical Modelling of Rods. *Handbook of Numerical Analysis*, Vol. IV, (P. G. Ciarlet, J. L. Lions), North-Holland, Amsterdam, (1996), pp. 487–974.

A new numerical scheme for Von-Mises plasticity with linear hardening

Lourenço Beirão da Veiga[1]

Dipartimento di Matematica
Università di Pavia
Via Ferrata 1, 27100 Pavia, Italy

Fundação para a Ciência e a Tecnologia
Lisbon, Portugal

Abstract

The problem addressed will be that of solving constitutive laws for plastic materials. The necessity to implement this local stress-strain relation on every quadrature node of the finite element grid used calls for a method which is quick, robust and sufficiently accurate.

Considering the classical constitutive law of von-Mises associative plasticity with linear isotropic and kinematic hardening, I will present a new integration scheme based on the computation of an integration factor ([1]). This method is consistent with the Yield surface condition (which is a "most wanted" property for numerical schemes in plasticity) and is exact whenever no isotropic hardening is present in the model.

The proposed scheme will be compared with the classical radial return map algorithm, a well estabilished and performing method ([2]). The comparision is based on various pointwise tests on different strain histories and an initial boundary value problem. Our method, which is still comparatively quick, clearly shows greater accuracy then the radial return map.

[1] F.Auricchio and L.Beirão da Veiga, *On a new integration scheme for von-Mises plasticity with linear hardening*, International Journal for Numerical Methods in Engineering 56:1375-1396 (2003)

[2] J.C.Simo and T.J.R.Hughes, *Computational Inelasticity*, Springer-Verlag New York (1998)

[1] Work in collaboration with F.Auricchio from the Department of Structural Mechanics of Pavia