Finite Element Methods for Non-Fickian Diffusion in Viscoelastic Polymers

A thesis submitted for the degree of
Doctor of Philosophy

by
Norbert Bauermeister

School of Information Systems, Computing and Mathematics
Brunel University

December 2007

Brunel University, Uxbridge
School of Information Systems,
Computing and Mathematics

Norbert Bauermeister
PhD thesis
December 2007

Finite Element Methods for Non-Fickian Diffusion in Viscoelastic Polymers

Abstract

The nonlinear problem of non-Fickian polymer diffusion is considered. The model is taken from Cohen, White & Witelski, SIAM J. Appl. Math. 55, pp. 348–368, 1995. It consists of a diffusion equation with an adjoined spatially local evolution equation for a viscoelastic stress. In their simplest form the equations are

$$\begin{aligned} u_t - \nabla^2 u - \nabla^2 \sigma &= 0, \\ \sigma_t + \gamma(u)\sigma &= u, \end{aligned}$$

for certain bounded functions $\gamma : \mathbb{R} \to \mathbb{R}$.

Numerical schemes are presented that are based, spatially, on the Galerkin finite element method and, temporally, on the Crank-Nicolson method. Special attention is paid to linearising the discrete equations by extrapolating the value of the discrete solution from previous timesteps. Optimal a priori error estimates are given, based on the assumption that the exact solution and the domain possess certain regularity properties. The error estimate is, in the case of approximation by piecewise linear polynomials,

$$\begin{aligned} &\max_n \big(\|u(t_n) - u_n^h\|_0 + \|\sigma(t_n) - \sigma_n^h\|_0 \big) \\ &\quad + h \Big(\sum_n \|\bar{u}_n + \bar{\sigma}_n - \bar{u}_n^h - \bar{\sigma}_n^h\|_1^2 \Big)^{1/2} \leq C(k^2 + h^2), \end{aligned}$$

where $\bar{v}_n = \frac{1}{2}(v(t_n) + v(t_{n-1}))$ for $v \in \{u, \sigma, u^h, \sigma^h\}$. Numerical experiments are given to support the error estimates.

For the variational formulation that forms the basis of the discrete schemes the existence and uniqueness of solutions in one space dimension is established.

Further numerical experiments are included to show some of the behaviour of the solutions of the model.

Contents

List of Tables **6**

List of Figures **9**

List of Theorems **11**

Acknowledgements **13**

1 Introduction **14**
1.1 Overview of the chapters 14

2 Mathematical Background **16**
2.1 Elliptic problems 17
2.1.1 The domain Ω 17
2.1.2 Sobolev spaces and their embeddings 17
2.1.3 Laplace equation 20
2.1.4 Inverse Laplacian 22
2.2 Time dependent problems 22
2.2.1 Spaces for time dependent problems 23
2.2.2 Unique solution of ODEs 25
2.2.3 Linear parabolic equations 25
2.3 Space discretisation 26
2.3.1 Mesh and finite element spaces 26
2.3.2 Inverse estimates 28
2.3.3 Interpolation 28
2.3.4 L_2-projection and elliptic projection 30

3 The Cohen Model for Non-Fickian Diffusion **33**
3.1 The Cohen model 33
3.2 Physical interpretation of the model 34
3.3 Notes on literature 35
3.4 Rescaling the equations 36
3.5 The nonlinearity 37
3.5.1 Choices for nonlinearity 37
3.5.2 Basic properties of γ 40

4 Existence and Uniqueness **42**
4.1 Variational formulation 42
4.1.1 Two approaches for getting results 43
4.2 Boundedness of exact solutions 44
4.2.1 Using the inverse Laplacian 44

4.2.2 Using the $u + \sigma$-approach . . . 45
4.3 Unique solution of linearised problem . . . 48
4.3.1 Weighted norms . . . 48
4.3.2 The result . . . 50
4.4 Unique solution in 1D . . . 52
4.4.1 An iteration for the nonlinear problem . . . 52
4.4.2 The fixed point of the iteration . . . 56
4.4.3 What about $d = 2$ and $d = 3$? . . . 58

5 Discretisation and Numerical Scheme 59
5.1 Time discretisation . . . 59
5.1.1 Notation . . . 59
5.1.2 Taylor estimates . . . 60
5.1.3 Discrete inverse Laplacian . . . 61
5.2 Discrete formulation . . . 62
5.3 Matrix formulation . . . 63
5.3.1 Matrix form for $Q = 3$ . . . 66
5.4 About uniqueness of discrete solutions . . . 67
5.4.1 Methods $Q \in \{1, 2\}$ . . . 67
5.4.2 Method $Q = 3$ . . . 68
5.4.3 Methods $Q \in \{4, 5\}$ . . . 68
5.5 Boundedness and stability of discrete solution . . . 71
5.5.1 Bound on solution . . . 71
5.5.2 Bound on differences . . . 77

6 Error Estimate 79
6.1 Preparations for the error estimate . . . 79
6.2 Error inequalities . . . 81
6.3 Nonlinearity errors . . . 84
6.4 Error estimate . . . 94
6.4.1 Why so complicated? . . . 98
6.5 About method $Q = 3$ . . . 99

7 Numerical Results 101
7.1 Another discrete formulation . . . 101
7.2 Implementation . . . 102
7.2.1 Defining errors . . . 103
7.3 Testing the programs for exactness . . . 103
7.3.1 Degree 4 polynomials in 3D . . . 104
7.4 Convergence tables . . . 104
7.4.1 Overview of tables . . . 105
7.4.2 Influence of time step k for $Q \in \{4, 5\}$ . . . 106
7.4.3 Summary of the tables . . . 106
7.5 Behaviour of the solution . . . 144
7.5.1 Sharp front in 1D . . . 144
7.5.2 Infinite domain in 1D: $\Omega = (0, \infty)$ . . . 144
7.5.3 Wave-like solution in 1D . . . 158
7.5.4 Sharp front in 2D . . . 160
7.5.5 Ring-shaped domain in 2D . . . 167
7.5.6 Sharp front in 3D . . . 170

8 Conclusion **173**
8.1 Suggestions for further work 173

A More on Mathematical Background **175**
A.1 Basic inequalities 175
A.2 Banach's fixed point theorem 176
A.3 Bounds for norms in finite-dimensional spaces 177
A.4 Gelfand triple 180
A.5 Positive definite matrices 183

B Amann's Result of Existence and Uniqueness **185**

References **188**

List of Symbols **191**

Index **196**

List of Tables

7.1 k-convergence for $\sigma_D = 1$, γ_3, $Q = 1$, $d = 2$, $r = 1$: e_{comb} 108
7.2 k-convergence for $\sigma_D = 1$, γ_3, $Q = 1$, $d = 2$, $r = 1$: e_0, e_1 108
7.3 k-convergence for $\sigma_D = 1$, γ_3, $Q = 2$, $d = 2$, $r = 1$: e_{comb} 108
7.4 k-convergence for $\sigma_D = 1$, γ_3, $Q = 2$, $d = 2$, $r = 1$: e_0, e_1 109
7.5 k-convergence for $\sigma_D = 1$, γ_3, $Q = 3$, $d = 2$, $r = 1$: e_{comb} 109
7.6 k-convergence for $\sigma_D = 1$, γ_3, $Q = 3$, $d = 2$, $r = 1$: e_0, e_1 109
7.7 k-convergence for $\sigma_D = 1$, γ_3, $Q = 4$, $d = 2$, $r = 1$: e_{comb} 110
7.8 k-convergence for $\sigma_D = 1$, γ_3, $Q = 4$, $d = 2$, $r = 1$: e_0, e_1 110
7.9 k-convergence for $\sigma_D = 1$, γ_3, $Q = 5$, $d = 2$, $r = 1$: e_{comb} 110
7.10 k-convergence for $\sigma_D = 1$, γ_3, $Q = 5$, $d = 2$, $r = 1$: e_0, e_1 111
7.11 k-convergence for $\sigma_D = 1$, γ_1, $Q = 1$, $d = 2$, $r = 1$: e_{comb} 111
7.12 k-convergence for $\sigma_D = 1$, γ_1, $Q = 1$, $d = 2$, $r = 1$: e_0, e_1 111
7.13 k-convergence for $\sigma_D = 1$, γ_1, $Q = 2$, $d = 2$, $r = 1$: e_{comb} 112
7.14 k-convergence for $\sigma_D = 1$, γ_1, $Q = 2$, $d = 2$, $r = 1$: e_0, e_1 112
7.15 k-convergence for $\sigma_D = 1$, γ_1, $Q = 3$, $d = 2$, $r = 1$: e_{comb} 112
7.16 k-convergence for $\sigma_D = 1$, γ_1, $Q = 3$, $d = 2$, $r = 1$: e_0, e_1 113
7.17 k-convergence for $\sigma_D = 1$, γ_1, $Q = 4$, $d = 2$, $r = 1$: e_{comb} 113
7.18 k-convergence for $\sigma_D = 1$, γ_1, $Q = 4$, $d = 2$, $r = 1$: e_0, e_1 113
7.19 k-convergence for $\sigma_D = 1$, γ_1, $Q = 5$, $d = 2$, $r = 1$: e_{comb} 114
7.20 k-convergence for $\sigma_D = 1$, γ_1, $Q = 5$, $d = 2$, $r = 1$: e_0, e_1 114
7.21 k-convergence for $\sigma_D = 0$, γ_3, $Q = 1$, $d = 2$, $r = 1$: e_{comb} 114
7.22 k-convergence for $\sigma_D = 0$, γ_3, $Q = 1$, $d = 2$, $r = 1$: e_0, e_1 115
7.23 k-convergence for $\sigma_D = 0$, γ_3, $Q = 2$, $d = 2$, $r = 1$: e_{comb} 115
7.24 k-convergence for $\sigma_D = 0$, γ_3, $Q = 2$, $d = 2$, $r = 1$: e_0, e_1 115
7.25 k-convergence for $\sigma_D = 0$, γ_3, $Q = 3$, $d = 2$, $r = 1$: e_{comb} 116
7.26 k-convergence for $\sigma_D = 0$, γ_3, $Q = 3$, $d = 2$, $r = 1$: e_0, e_1 116
7.27 k-convergence for $\sigma_D = 0$, γ_3, $Q = 4$, $d = 2$, $r = 1$: e_{comb} 116
7.28 k-convergence for $\sigma_D = 0$, γ_3, $Q = 4$, $d = 2$, $r = 1$: e_0, e_1 117
7.29 k-convergence for $\sigma_D = 0$, γ_3, $Q = 5$, $d = 2$, $r = 1$: e_{comb} 117
7.30 k-convergence for $\sigma_D = 0$, γ_3, $Q = 5$, $d = 2$, $r = 1$: e_0, e_1 117
7.31 k-convergence for $\sigma_D = 0$, γ_1, $Q = 1$, $d = 2$, $r = 1$: e_{comb} 118
7.32 k-convergence for $\sigma_D = 0$, γ_1, $Q = 1$, $d = 2$, $r = 1$: e_0, e_1 118
7.33 k-convergence for $\sigma_D = 0$, γ_1, $Q = 2$, $d = 2$, $r = 1$: e_{comb} 118
7.34 k-convergence for $\sigma_D = 0$, γ_1, $Q = 2$, $d = 2$, $r = 1$: e_0, e_1 119
7.35 k-convergence for $\sigma_D = 0$, γ_1, $Q = 3$, $d = 2$, $r = 1$: e_{comb} 119
7.36 k-convergence for $\sigma_D = 0$, γ_1, $Q = 3$, $d = 2$, $r = 1$: e_0, e_1 119
7.37 k-convergence for $\sigma_D = 0$, γ_1, $Q = 4$, $d = 2$, $r = 1$: e_{comb} 120
7.38 k-convergence for $\sigma_D = 0$, γ_1, $Q = 4$, $d = 2$, $r = 1$: e_0, e_1 120
7.39 k-convergence for $\sigma_D = 0$, γ_1, $Q = 5$, $d = 2$, $r = 1$: e_{comb} 120
7.40 k-convergence for $\sigma_D = 0$, γ_1, $Q = 5$, $d = 2$, $r = 1$: e_0, e_1 121

7.41 k-convergence for $\sigma_D = 1$, γ_3, $Q = 1$, $d = 1$, $r = 1$: e_{comb} 121
7.42 k-convergence for $\sigma_D = 1$, γ_3, $Q = 1$, $d = 1$, $r = 1$: e_0, e_1 121
7.43 k-convergence for $\sigma_D = 1$, γ_3, $Q = 1$, $d = 3$, $r = 1$: e_{comb} 122
7.44 k-convergence for $\sigma_D = 1$, γ_3, $Q = 1$, $d = 3$, $r = 1$: e_0, e_1 122
7.45 h-convergence for $\sigma_D = 1$, γ_3, $Q = 1$, $d = 2$, $r = 1$, $k = 1.0$: e_{comb} . . . 122
7.46 h-convergence for $\sigma_D = 1$, γ_3, $Q = 1$, $d = 2$, $r = 1$, $k = 1.0$: e_0, e_1 . . . 123
7.47 h-convergence for $\sigma_D = 1$, γ_3, $Q = 2$, $d = 2$, $r = 1$, $k = 1.0$: e_{comb} . . . 123
7.48 h-convergence for $\sigma_D = 1$, γ_3, $Q = 2$, $d = 2$, $r = 1$, $k = 1.0$: e_0, e_1 . . . 123
7.49 h-convergence for $\sigma_D = 1$, γ_3, $Q = 3$, $d = 2$, $r = 1$, $k = 1.0$: e_{comb} . . . 124
7.50 h-convergence for $\sigma_D = 1$, γ_3, $Q = 3$, $d = 2$, $r = 1$, $k = 1.0$: e_0, e_1 . . . 124
7.51 h-convergence for $\sigma_D = 1$, γ_3, $Q = 4$, $d = 2$, $r = 1$, $k = 0.0001$: e_{comb} . . 124
7.52 h-convergence for $\sigma_D = 1$, γ_3, $Q = 4$, $d = 2$, $r = 1$, $k = 0.0001$: e_0, e_1 . 125
7.53 h-convergence for $\sigma_D = 1$, γ_3, $Q = 5$, $d = 2$, $r = 1$, $k = 0.0001$: e_{comb} . . 125
7.54 h-convergence for $\sigma_D = 1$, γ_3, $Q = 5$, $d = 2$, $r = 1$, $k = 0.0001$: e_0, e_1 . 125
7.55 h-convergence for $\sigma_D = 1$, γ_1, $Q = 1$, $d = 2$, $r = 1$, $k = 1.0$: e_{comb} . . . 126
7.56 h-convergence for $\sigma_D = 1$, γ_1, $Q = 1$, $d = 2$, $r = 1$, $k = 1.0$: e_0, e_1 . . . 126
7.57 h-convergence for $\sigma_D = 1$, γ_1, $Q = 2$, $d = 2$, $r = 1$, $k = 1.0$: e_{comb} . . . 126
7.58 h-convergence for $\sigma_D = 1$, γ_1, $Q = 2$, $d = 2$, $r = 1$, $k = 1.0$: e_0, e_1 . . . 127
7.59 h-convergence for $\sigma_D = 1$, γ_1, $Q = 3$, $d = 2$, $r = 1$, $k = 1.0$: e_{comb} . . . 127
7.60 h-convergence for $\sigma_D = 1$, γ_1, $Q = 3$, $d = 2$, $r = 1$, $k = 1.0$: e_0, e_1 . . . 127
7.61 h-convergence for $\sigma_D = 1$, γ_1, $Q = 4$, $d = 2$, $r = 1$, $k = 0.0001$: e_{comb} . . 128
7.62 h-convergence for $\sigma_D = 1$, γ_1, $Q = 4$, $d = 2$, $r = 1$, $k = 0.0001$: e_0, e_1 . 128
7.63 h-convergence for $\sigma_D = 1$, γ_1, $Q = 5$, $d = 2$, $r = 1$, $k = 0.0001$: e_{comb} . . 128
7.64 h-convergence for $\sigma_D = 1$, γ_1, $Q = 5$, $d = 2$, $r = 1$, $k = 0.0001$: e_0, e_1 . 128
7.65 h-convergence for $\sigma_D = 0$, γ_3, $Q = 1$, $d = 2$, $r = 1$, $k = 1.0$: e_{comb} . . . 129
7.66 h-convergence for $\sigma_D = 0$, γ_3, $Q = 1$, $d = 2$, $r = 1$, $k = 1.0$: e_0, e_1 . . . 129
7.67 h-convergence for $\sigma_D = 0$, γ_3, $Q = 2$, $d = 2$, $r = 1$, $k = 1.0$: e_{comb} . . . 129
7.68 h-convergence for $\sigma_D = 0$, γ_3, $Q = 2$, $d = 2$, $r = 1$, $k = 1.0$: e_0, e_1 . . . 130
7.69 h-convergence for $\sigma_D = 0$, γ_3, $Q = 3$, $d = 2$, $r = 1$, $k = 1.0$: e_{comb} . . . 130
7.70 h-convergence for $\sigma_D = 0$, γ_3, $Q = 3$, $d = 2$, $r = 1$, $k = 1.0$: e_0, e_1 . . . 130
7.71 h-convergence for $\sigma_D = 0$, γ_3, $Q = 4$, $d = 2$, $r = 1$, $k = 0.0001$: e_{comb} . . 131
7.72 h-convergence for $\sigma_D = 0$, γ_3, $Q = 4$, $d = 2$, $r = 1$, $k = 0.0001$: e_0, e_1 . 131
7.73 h-convergence for $\sigma_D = 0$, γ_3, $Q = 5$, $d = 2$, $r = 1$, $k = 0.0001$: e_{comb} . . 132
7.74 h-convergence for $\sigma_D = 0$, γ_3, $Q = 5$, $d = 2$, $r = 1$, $k = 0.0001$: e_0, e_1 . 132
7.75 h-convergence for $\sigma_D = 0$, γ_1, $Q = 1$, $d = 2$, $r = 1$, $k = 1.0$: e_{comb} . . . 132
7.76 h-convergence for $\sigma_D = 0$, γ_1, $Q = 1$, $d = 2$, $r = 1$, $k = 1.0$: e_0, e_1 . . . 133
7.77 h-convergence for $\sigma_D = 0$, γ_1, $Q = 2$, $d = 2$, $r = 1$, $k = 1.0$: e_{comb} . . . 133
7.78 h-convergence for $\sigma_D = 0$, γ_1, $Q = 2$, $d = 2$, $r = 1$, $k = 1.0$: e_0, e_1 . . . 133
7.79 h-convergence for $\sigma_D = 0$, γ_1, $Q = 3$, $d = 2$, $r = 1$, $k = 1.0$: e_{comb} . . . 134
7.80 h-convergence for $\sigma_D = 0$, γ_1, $Q = 3$, $d = 2$, $r = 1$, $k = 1.0$: e_0, e_1 . . . 134
7.81 h-convergence for $\sigma_D = 0$, γ_1, $Q = 4$, $d = 2$, $r = 1$, $k = 0.0001$: e_{comb} . . 134
7.82 h-convergence for $\sigma_D = 0$, γ_1, $Q = 4$, $d = 2$, $r = 1$, $k = 0.0001$: e_0, e_1 . 135
7.83 h-convergence for $\sigma_D = 0$, γ_1, $Q = 5$, $d = 2$, $r = 1$, $k = 0.0001$: e_{comb} . . 135
7.84 h-convergence for $\sigma_D = 0$, γ_1, $Q = 5$, $d = 2$, $r = 1$, $k = 0.0001$: e_0, e_1 . 135
7.85 h-convergence for $\sigma_D = 1$, γ_3, $Q = 1$, $d = 1$, $r = 1$, $k = 1.0$: e_{comb} . . . 136
7.86 h-convergence for $\sigma_D = 1$, γ_3, $Q = 1$, $d = 1$, $r = 1$, $k = 1.0$: e_0, e_1 . . . 136
7.87 h-convergence for $\sigma_D = 1$, γ_3, $Q = 1$, $d = 3$, $r = 1$, $k = 1.0$: e_{comb} . . . 137
7.88 h-convergence for $\sigma_D = 1$, γ_3, $Q = 1$, $d = 3$, $r = 1$, $k = 1.0$: e_0, e_1 . . . 137
7.89 h-convergence for $\sigma_D = 1$, γ_3, $Q = 3$, $d = 1$, $r = 2$, $k = 1.0$: e_{comb} . . . 137
7.90 h-convergence for $\sigma_D = 1$, γ_3, $Q = 3$, $d = 1$, $r = 2$, $k = 1.0$: e_0, e_1 . . . 138
7.91 h-convergence for $\sigma_D = 1$, γ_3, $Q = 3$, $d = 1$, $r = 3$, $k = 1.0$: e_{comb} . . . 139

7.92 h-convergence for $\sigma_D = 1$, γ_3, $Q = 3$, $d = 1$, $r = 3$, $k = 1.0$: e_0, e_1 . . . 139
7.93 h-convergence for $\sigma_D = 1$, γ_3, $Q = 3$, $d = 1$, $r = 4$, $k = 1.0$: e_{comb} . . . 140
7.94 h-convergence for $\sigma_D = 1$, γ_3, $Q = 3$, $d = 1$, $r = 4$, $k = 1.0$: e_0, e_1 . . . 140
7.95 h-convergence for $\sigma_D = 1$, γ_3, $Q = 4$, $d = 1$, $r = 1$, $k = 0.0001$: e_{comb} . . 140
7.96 h-convergence for $\sigma_D = 1$, γ_3, $Q = 4$, $d = 1$, $r = 1$, $k = 0.0001$: e_0, e_1 . 141
7.97 h-convergence for $\sigma_D = 1$, γ_3, $Q = 4$, $d = 1$, $r = 1$, $k = 0.25$: e_{comb} . . . 141
7.98 h-convergence for $\sigma_D = 1$, γ_3, $Q = 4$, $d = 1$, $r = 1$, $k = 0.25$: e_0, e_1 . . . 141
7.99 h-convergence for $\sigma_D = 1$, γ_3, $Q = 5$, $d = 1$, $r = 1$, $k = 0.0001$: e_{comb} . . 142
7.100h-convergence for $\sigma_D = 1$, γ_3, $Q = 5$, $d = 1$, $r = 1$, $k = 0.0001$: e_0, e_1 . 142
7.101h-convergence for $\sigma_D = 1$, γ_3, $Q = 5$, $d = 1$, $r = 1$, $k = 0.25$: e_{comb} . . . 142
7.102h-convergence for $\sigma_D = 1$, γ_3, $Q = 5$, $d = 1$, $r = 1$, $k = 0.25$: e_0, e_1 . . . 143
7.103k-convergence for $\sigma_D = 1$, γ_3, $Q = 1$, $d = 2$, $r = 1$, $\hat{\gamma} = 10$: e_{comb} 143
7.104k-convergence for $\sigma_D = 1$, γ_3, $Q = 1$, $d = 2$, $r = 1$, $\hat{\gamma} = 10$: e_0, e_1 143
7.105Infinite domain in 1D: Summary of types 148

List of Figures

2.1 Function v_ϵ and its interpolant $I^h v_\epsilon$ from Example 2.30. 29

3.1 Graphs of $\gamma = \gamma_1$ and its derivatives. This is the tanh-version of γ. . . . 37
3.2 Graphs of $\gamma = \gamma_2$ and its derivatives. This is the polynomial version of γ. 39
3.3 Graphs of $\gamma = \gamma_3$ and its derivatives. This is the $\sin(x) + x$-version of γ. 40

5.1 A non-unique discrete solution for $Q = 4$ 70
5.2 Discrete solution for $Q = 4$ now unique 70
5.3 A non-unique discrete solution for $Q = 5$ 70
5.4 Discrete solution for $Q = 5$ now unique 70

7.1 Numerical solutions $u^h(t = 1)$ for $Q = 4$ and $h = 1/16$ 107
7.2 Numerical solution on $\Omega = (0, 1)$ 145
7.3 Typical shape of functions G and D for Type I. 147
7.4 Typical shape of functions G and D for Type II. 148
7.5 Typical shape of functions G and D for Type III. 149
7.6 Type I: $\gamma(u)$, $\gamma'(u)$, $\gamma''(u)$, $G(u)$ and $D(u)$ 151
7.7 Type I: $u(x)$, $\sigma(x)$, $(u + \sigma)(x)$ 152
7.8 Type II: $\gamma(u)$, $\gamma'(u)$, $\gamma''(u)$, $G(u)$ and $D(u)$ 153
7.9 Type II: Zoom to $G(u)$ and $D(u)$ 153
7.10 Type IIa: $u(x)$, $\sigma(x)$, $(u + \sigma)(x)$ 154
7.11 Type IId: $u(x)$, $\sigma(x)$, $(u + \sigma)(x)$, $x_m(t)$ 154
7.12 Type III: $\gamma(u)$, $\gamma'(u)$, $\gamma''(u)$, $G(u)$ and $D(u)$ 155
7.13 Type III: Zoom to $G(u)$ and $D(u)$ 155
7.14 Type IIIa: $u(x)$, $\sigma(x)$, $(u + \sigma)(x)$ 156
7.15 Type IIIb: $u(x)$, $\sigma(x)$, $(u + \sigma)(x)$, $x_m(t)$ 156
7.16 Type IIIc: $u(x)$, $\sigma(x)$, $(u + \sigma)(x)$, $x_m(t)$ 157
7.17 Type IIId: $u(x)$, $\sigma(x)$, $(u + \sigma)(x)$, $x_m(t)$ 157
7.18 Wave-like numerical solution 159
7.19 Colour map for the following plots 160
7.20 Domain for 2D example 160
7.21 Sharp front in 2D: Numerical solution u at $t = 1.8$ and $t = 3.6$ 161
7.22 Sharp front in 2D: Numerical solution u at $t = 15$ and $t = 30$ 162
7.23 Sharp front in 2D: Numerical solution σ at $t = 1.8$ and $t = 3.6$ 163
7.24 Sharp front in 2D: Numerical solution σ at $t = 15$ and $t = 30$ 164
7.25 Sharp front in 2D: Numerical solution $u + \sigma$ at $t = 1.8$ and $t = 3.6$ . . . 165
7.26 Sharp front in 2D: Numerical solution $u + \sigma$ at $t = 15$ and $t = 30$ 166
7.27 Ring-shaped domain in 2D 167
7.28 Ring-shaped domain in 2D: Numerical solution u and σ at $t = 20$ 168
7.29 Ring-shaped domain in 2D: Numerical solution $u + \sigma$ at $t = 20$. 169
7.30 Domain for 3D example 170

7.31 Sharp front in 3D: Numerical solution u and σ at $t = 10$ 171
7.32 Sharp front in 3D: Numerical solution $u + \sigma$ at $t = 10$. 172

List of Theorems

Theorem 2.4 (Sobolev embedding) . . . 19
Lemma 2.6 (Hilbert space V) . . . 20
Lemma 2.7 (Gelfand triple) . . . 21
Lemma 2.8 (Green's formula) . . . 21
Theorem 2.10 (Unique solution of Laplace equation) . . . 22
Lemma 2.11 (Inverse Laplacian) . . . 22
Lemma 2.12 (The space $C^m([a,b];X)$) . . . 23
Lemma 2.13 (An equality for $(\dot{v},v)_0$) . . . 23
Lemma 2.14 (The space $W_p^m(a,b;X)$) . . . 23
Lemma 2.15 (Embeddings of time-dependent function spaces) . . . 24
Lemma 2.17 (Properties of $\mathcal{W}(0,T;V,V')$) . . . 24
Proposition 2.18 (Unique solution for ODEs) . . . 25
Theorem 2.19 (Existence and uniqueness for linear parabolic problem) . . . 25
Lemma 2.24 (Lagrange basis functions) . . . 27
Theorem 2.27 (Inverse Estimates) . . . 28
Lemma 2.29 (Approximation error of interpolation) . . . 28
Lemma 2.31 (Scott-Zhang interpolation) . . . 30
Lemma 2.33 (Stability of L_2- and elliptic projections) . . . 30
Lemma 2.34 (Approximation error for L_2- and elliptic projections) . . . 31

Lemma 3.5 (Rescaling the Cohen model) . . . 36
Lemma 3.6 (Usual choice for γ) . . . 37
Lemma 3.7 (Another choice for γ) . . . 38
Lemma 3.8 (A third choice for γ) . . . 38
Lemma 3.9 (Difference of γ-values in L_p-norm) . . . 40
Lemma 3.10 (A bound for C^2-functions) . . . 40
Lemma 3.11 (Another bound for γ-values) . . . 41

Lemma 4.4 (Consequences of triangle inequality) . . . 44
Proposition 4.5 (Basic stability in weak norms) . . . 44
Proposition 4.6 (Basic stability in stronger norms) . . . 46
Lemma 4.8 (Equivalence of weighted max-norms) . . . 48
Lemma 4.9 (Trick with weighted max-norm) . . . 49
Proposition 4.10 (Unique solution of linearised Cohen model) . . . 50
Lemma 4.12 (An iteration for the nonlinear problem) . . . 52
Lemma 4.14 (Bound for difference of iterates) . . . 53
Lemma 4.16 ($\|\sigma\|_{L_\infty(\Omega)}$-bound of iterate in 1D) . . . 55
Lemma 4.17 (Banach fixed point theorem for subsets) . . . 56
Proposition 4.18 (Unique solution in 1D) . . . 57

Lemma 5.1 (Taylor estimates) . . . 60

Lemma 5.3 (Properties of discrete inverse Laplacian) 62
Lemma 5.7 (Matrix form of discretised problem) 64
Lemma 5.8 (Alternative matrix formulation for $Q = 3$) 66
Lemma 5.9 (Unique discrete solution for $Q = 1, 2$) 67
Lemma 5.12 (Upper bound for $\mathscr{B}_n^Q$) 71
Lemma 5.13 (Lower bounds for $\mathscr{B}_n^Q$) 71
Proposition 5.14 (Boundedness of discrete solutions using inverse Laplacian) . 72
Proposition 5.16 (Boundedness of discrete solutions using $u + \sigma$-terms) 74
Corollary 5.17 (Boundedness of discrete solutions) 76
Proposition 5.18 (Bound on differences) 77

Lemma 6.1 (Approximation error for elliptic projection) 79
Lemma 6.2 (Bounds involving ξ_n and θ_n) 80
Lemma 6.3 (Error inequalities) 81
Lemma 6.4 (Intermediate nonlinearity error for methods $Q \in \{2, 5\}$) 85
Lemma 6.5 (Intermediate nonlinearity error for methods $Q \in \{1, 4\}$) 87
Lemma 6.6 (Extrapolation errors) 88
Lemma 6.7 (Nonlinearity errors for methods $Q \in \{1, 2, 4, 5\}$) 92
Theorem 6.8 (An a priori error estimate) 94
Lemma 6.9 (Tentative nonlinearity error for $Q = 3$) 99

Lemma A.1 (Jensen's inequality) 175
Lemma A.2 (Continuous Grönwall lemma) 175
Lemma A.3 (Continuous Grönwall lemma, second form) 176
Lemma A.4 (Discrete Grönwall lemma) 176
Theorem A.5 (Banach's fixed point theorem) 177
Proposition A.8 (Criterion for convexity/concavity) 177
Corollary A.9 (x^p is convex if $p \geq 1$) 177
Proposition A.10 (Sub- and superadditivity) 177
Corollary A.12 (x^p is superadditive) 178
Lemma A.13 (Generalising the property of convex functions) 179
Proposition A.14 (Equivalence of p-norms in $\mathbb{R}^n$) 179
Theorem A.16 (Hahn-Banach) 181
Lemma A.17 (An equality with the norm of the dual space) 181
Lemma A.21 (Properties of dual mapping) 182
Lemma A.22 (Injectivity of dual mapping) 182
Lemma A.23 (Embedding of dual spaces) 182
Theorem A.24 (Riesz' representation theorem) 182
Proposition A.26 (Gelfand triple) 183
Lemma A.30 (A counter-example) 184
Lemma A.31 (Product of symmetric positive definite matrices) 184

Corollary B.1 (Amann's result specialised for the Cohen model) 185
Theorem B.2 (Amann) 186

Acknowledgements

I would like to take this opportunity to thank the people who have helped me in many different ways to finish this thesis.

First of all, I thank my supervisor Simon Shaw, without whom this thesis would never have been possible. He always had useful advice and the work with him has been fun.

I am very grateful to Manu for her invaluable support, especially during the final months of writing. The help from my family also encouraged me a lot.

I would like to express my warmest thanks to Dave Needham, John Leach and John Merkin, for valuable discussions and a better understanding of the equations, and Andrew Wathen for pointing out a connection to product approximations.

This work relies heavily on open source software and I am indebted to the developers of the many free software packages like GCC, Gnuplot, LaTeX, Python, and many more. Special thanks goes to the authors of ALBERTA finite element toolbox.

Last but not least, I would also like to thank Brunel University and EPSRC for their financial support.

Chapter 1

Introduction

The aim of this thesis is to give a standard finite element scheme for approximately solving a nonlinear PDE system, which was proposed in [CW91] and [CWW95] and which we will call the *Cohen model*. This model was formulated to explain experimental results for diffusion in viscoelastic polymers. It combines two departures from the framework of standard Fickian diffusion: firstly, a stress variable is added to give a viscoelastic memory effect and secondly, the evolution of this stress variable contains a nonlinearity modelling a phase change of the polymer. For a more detailed description, we refer the reader forward to Sections 3.2 and 3.3.

As a foundation for the results of this thesis, first a variational formulation of the model is developed and some partial results towards establishing the well-posedness of this formulation are given.

Then the finite element scheme is presented, giving a choice of five methods for discretising the nonlinearity of the system.

An a priori estimate is proved for four of the methods showing that the discrete solution will converge to the exact solution, provided the exact solution is sufficiently smooth. Numerical experiments to support the predicted convergence rates are given. This a priori estimate relies on a combination of two approaches to combine the evolution equations of the model. These approaches are introduced in Subsection 4.1.1. One of them uses the *inverse Laplacian* and the other works with the sum of concentration and stress. An explanation of why only the combination of both approaches seems to work to prove the estimate is given in Subsection 6.4.1.

Finally, the finite element scheme is used to demonstrate some of the behaviour of the Cohen model: The concentration can develop sharp fronts and wave-like behaviour.

1.1 Overview of the chapters

Chapter 2 presents some of the mathematical background from functional analysis and finite element theory which will be needed in the later chapters.

Chapter 3 states the problem that we will examine in this thesis, the Cohen model for non-Fickian diffusion. It also provides some background on the physical interpretation and on the history of this model.

Chapter 4 gives a variational formulation of the problem. This formulation is the basis for the finite element discretisation and error analysis. Here, we will look at questions of existence and uniqueness of the exact solution.

In Chapter 5 we present the discrete formulation of the problem and prove bounds for the discrete solution.

Chapter 6 is the central part of the thesis that contains the main result: the error estimate that shows the convergence of the discrete scheme.

Chapter 7 gives some numerical results: Tables of convergence confirm the convergence rates predicted by the error estimate. Plots of the numerical solution show how the solution behaves in some special cases.

Chapter 8 is the Conclusion, containing a review of the results of this thesis and suggestions for further work.

There are also two appendices: Appendix A contains more on the mathematical background and Appendix B gives an interpretation of a result by Amann on existence of solutions for the Cohen model.

To help the reader find particular items in this thesis, a List of Tables, a List of Figures and a List of Theorems is included at the beginning, after the Table of Contents. Also, the mathematical symbols that appear more than once can be looked up in the List of Symbols on page 191. Finally, there is an Index, starting on page 196.

Chapter 2

Mathematical Background

This chapter is intended as an introduction to the mathematical framework for the later chapters. The necessary background from functional analysis and finite element theory is presented, firstly to introduce the notation used in this thesis and secondly to state the results from standard theory that the reader can refer back to later.

Section 2.1 gives functional analytic tools that are used to deal with elliptic problems like the Laplace equation.

In Section 2.2 further results are given to extend the framework to time-dependent problems. An abstract existence and uniqueness result for parabolic problems is stated.

Section 2.3 presents a standard finite element discretisation of the Laplace equation. The same discretisation will be applied later to the main problem of this thesis.

Notation

The notation that is used in this thesis is fairly standard. Most symbols and other notation are defined where they are first needed. For a complete list of notation, we refer the reader to the List of Symbols on page 191 at the end of the thesis. Except for very basic mathematical notation, every item in that list has a reference to the page or equation in the main text where it is defined.

It is assumed that the reader is familiar with certain basic topics from functional analysis and linear algebra, like Banach spaces, Hilbert spaces, etc.

We use C to denote a generic constant, that may have different values in different places. When we want to define constants using explicit expressions we denote them by c_i for some $i = 1, 2, \ldots$. The value of *these* constants remains fixed until the end of that chapter.

A shorthand notation used frequently in abstract algebra is: If $f : X \to Y$ is a mapping from set X into set Y and $Z \subseteq X$ a subset, then we set

$$f(Z) := \{f(z) \ : \ z \in Z\}.$$

Another convention for this thesis is the use of the multiplication symbol '$\cdot$': It is used for normal multiplication and not for the Euclidean inner product. In the few cases, where we need the latter, it is denoted by $\boldsymbol{x}^T\boldsymbol{y}$ for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$.

Finally, for the notation of integrals, we frequently omit the measure terms such as $d\Omega$, $d\boldsymbol{x}$ from the integrals. For example, we write $\int_\Omega f$ instead of $\int_\Omega f d\Omega$.

2.1 Elliptic problems

2.1.1 The domain Ω

This subsection is concerned with the domain for the differential equations. It contains not only definitions but also assumptions that are made on the domain. Throughout this thesis, we only consider domains that satisfy the assumptions made below.

Certain results in this thesis require an even more regular domain. Since these restrictions will be quite strong, we will introduce a 'switch' in Definition 2.3. Using that definition, the regularity assumptions can be stated in a concise way. But first we start with the basic assumptions on the domain.

Definition 2.1 (Assumption on domain Ω).
Ω is a connected, open, bounded subset of $\mathbb{R}^d$, with dimension $d \in \{1, 2, 3\}$. Moreover Ω is assumed to have a regular boundary $\partial\Omega$ in the following sense: Either

- $\partial\Omega$ is C^∞-smooth, or
- Ω is a polyhedral domain, i.e. $\partial\Omega$ is a polyhedron, if $d = 3$, a polygon, if $d = 2$, or consists of two points, if $d = 1$.

In both cases, the *unit outward normal vector* $\boldsymbol{n}$ is defined almost everywhere on the boundary.

To prescribe different types of boundary condition when formulating the PDE system, we split the boundary of Ω into two parts:

Definition 2.2 (Dirichlet and Neumann boundaries).
The boundary of Ω, $\partial\Omega$, consists of two parts, Γ_D and Γ_N, such that

$$\Gamma_D \cup \Gamma_N = \partial\Omega, \quad \Gamma_D \cap \Gamma_N = \varnothing, \tag{2.1}$$

where Γ_D is closed and has strictly positive $(d-1)$-dimensional measure. Γ_D is called the Dirichlet boundary and Γ_N is called the Neumann boundary.

Now we come to the 'switch' that was mentioned at the beginning of this subsection.

Definition 2.3 ('$\mathcal{ER}^q$ holds').
With the assumptions from Definitions 2.1 and 2.2, we say that '$\mathcal{ER}^0$ holds'. Furthermore, we say that '$\mathcal{ER}^1$ holds' if two additional regularity assumptions are satisfied:

- If Ω is a polyhedral domain, then it is convex.
- The whole boundary is a Dirichlet boundary: $\partial\Omega = \Gamma_D$.

Note that $\mathcal{ER}^0$ is always assumed to hold without stating it explicitly. We will need the definition of $\mathcal{ER}^q$ later: If $\mathcal{ER}^1$ holds, the solution of the Laplace equation has a higher regularity (see Theorem 2.10) which is sometimes called *elliptic regularity*. Also, as a consequence, the L_2-error of the elliptic projection has a better bound, see Lemma 2.34. These better bounds will be used for the error estimate in Chapter 6 and we obtain a higher order of convergence for the discrete schemes.

2.1.2 Sobolev spaces and their embeddings

In this subsection several spaces of functions

$$v : \Omega \to \mathbb{R}$$

are defined. However, the Lebesgue and Sobolev spaces defined below are somewhat special. Strictly speaking their elements are not functions anymore, but equivalence

classes of functions. Two functions are identified when they differ only on a set of measure 0. For simplicity these equivalence classes are also called 'functions'. If v is a function from one of these spaces, and $\boldsymbol{x} \in \Omega$ is a point, then the expression $v(\boldsymbol{x})$ might not have a well-defined value. On the other hand, the Sobolev embedding theorem which is stated later in this subsection, asserts that for certain Sobolev spaces each function has a continuous function in its equivalence class. In these cases the pointwise function values $v(\boldsymbol{x})$ exist.

Defining the Sobolev spaces also involves derivatives of functions. In general these functions might not have classical derivatives, thus so-called *distributional derivatives* have to be used, see for example [DL88]. Since the calculation with the distributional derivatives follows the same rules as for the classical derivatives and since both derivatives coincide whenever the classical derivative exists, we do not distinguish between them.

In this subsection we state, along with the definitions, several continuous embeddings of the function spaces. For a definition of continuous embeddings, we refer the reader to the appendix, Definition A.18.

Space of Continuous Functions

First of all define $C^0(\bar{\Omega})$ as the space of continuous functions $\bar{\Omega} \to \mathbb{R}$. This space is a Banach space, when equipped with the norm

$$\|v\|_{C^0(\bar{\Omega})} := \max_{\boldsymbol{x} \in \bar{\Omega}} |v(\boldsymbol{x})|, \tag{2.2}$$

for any $v \in C^0(\bar{\Omega})$, see for example [BG87a, 2.3A]. Note that the maximum in (2.2) is attained, by the Weierstrass theorem, since we assume Ω to be bounded.

Lebesgue spaces

For any $1 \leq p \leq \infty$, the norm of the Lebesgue spaces is defined by

$$\begin{aligned} \|v\|_{L_p(\Omega)} &:= \left(\int_\Omega |v|^p \, d\Omega \right)^{1/p} \qquad \text{for } 1 \leq p < \infty, \\ \|v\|_{L_\infty(\Omega)} &:= \operatorname*{ess\,sup}_{\boldsymbol{x} \in \Omega} |v(\boldsymbol{x})|, \end{aligned} \tag{2.3}$$

and for any $1 \leq p \leq \infty$, $L_p(\Omega)$ is the set of Lebesgue-integrable functions for which the L_p-norm yields a finite value.

For $1 \leq p \leq q \leq \infty$ the Lebesgue spaces are continuously embedded as follows:

$$L_\infty(\Omega) \hookrightarrow L_q(\Omega) \hookrightarrow L_p(\Omega) \hookrightarrow L_1(\Omega).$$

Sobolev spaces

For $m \in \mathbb{N}$ and $1 \leq p \leq \infty$, the Sobolev space $W_p^m(\Omega)$ is defined by

$$W_p^m(\Omega) := \{ v \in L_p(\Omega) \ : \ \partial^\alpha v \in L_p(\Omega) \text{ if } |\alpha| \leq m \},$$

where $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d$ is a multi-index with $|\alpha| := \sum_{i=1}^d \alpha_i$ and $\partial^\alpha = \prod_{i=1}^d (\frac{\partial}{\partial x_i})^{\alpha_i}$. The derivatives used here are distributional derivatives. The norms for these spaces are

$$\begin{aligned} \|v\|_{W_p^m(\Omega)} &:= \Big(\sum_{|\alpha| \leq m} \|\partial^\alpha v\|_{L^p(\Omega)}^p \Big)^{1/p} \qquad \text{for } 1 \leq p < \infty, \\ \|v\|_{W_\infty^m(\Omega)} &:= \max_{|\alpha| \leq m} \|\partial^\alpha v\|_{L_\infty(\Omega)}. \end{aligned} \tag{2.4}$$

$W_p^m(\Omega)$ together with $\|\cdot\|_{W_p^m(\Omega)}$ is a Banach space. Note that $W_p^0(\Omega) = L_p(\Omega)$.

Sobolev Hilbert spaces

Set, for $m \in \mathbb{N}$,

$$H^m(\Omega) := W_2^m(\Omega),$$

and define the inner product

$$(v, w)_{H^m(\Omega)} := \sum_{|\alpha| \leq m} \int_\Omega \partial^\alpha v \cdot \partial^\alpha w \, d\Omega,$$

then $H^m(\Omega)$ together with $(\cdot, \cdot)_{H^m(\Omega)}$ is a Hilbert space. Note that $H^0(\Omega) = L_2(\Omega)$. As for any Hilbert space, the inner product satisfies the Cauchy-Schwarz inequality:

$$(v, w)_{H^m(\Omega)} \leq \|v\|_{H^m(\Omega)} \cdot \|w\|_{H^m(\Omega)} \quad \text{for } v, w \in H^m(\Omega). \tag{2.5}$$

To have a shorter notation for the L_2-norm and inner product, we set

$$\begin{aligned} (\cdot, \cdot)_0 &:= (\cdot, \cdot)_{L_2(\Omega)}, \\ \|\cdot\|_0 &:= \|\cdot\|_{L_2(\Omega)}. \end{aligned} \tag{2.6}$$

This notation is used for example in [Bra97].

The Product space $L_2(\Omega)^2$

We will sometimes use the space $L_2(\Omega)^2 = L_2(\Omega) \times L_2(\Omega)$ with norm

$$\|(v, w)\|_{L_2(\Omega)^2} := \left(\|v\|^2_{L_2(\Omega)} + \|w\|^2_{L_2(\Omega)} \right)^{1/2}, \tag{2.7}$$

for any $v, w \in L_2(\Omega)$. As $L_2(\Omega)$ is a Hilbert space it follows that $L_2(\Omega)^2$ is also a Hilbert space, see also [BG87b, Example 1.15].

Sobolev embeddings

As was mentioned before, Sobolev spaces consist of equivalence classes of functions. However, the elements of certain Sobolev spaces can be identified with continuous functions and the Sobolev embedding theorem gives conditions under which this is the case. We cite a version of the Sobolev embedding theorem from [BS94]:

Theorem 2.4 (Sobolev embedding).
Let k be a positive integer, m an integer with $m < k$ and let p be a real number in the range $1 \leq p < \infty$ such that

$$\begin{aligned} k - m &\geq d \quad \text{when} \quad p = 1, \\ k - m &> d/p \quad \text{when} \quad p > 1, \end{aligned}$$

where d is the dimension of Ω. Then there is a constant C such that for all $u \in W_p^k(\Omega)$

$$\|u\|_{W_\infty^m(\Omega)} \leq C \|u\|_{W_p^k(\Omega)}.$$

Moreover, there is an m times continuously differentiable function in the $L_\infty(\Omega)$ equivalence class of u.

Proof. This is [BS94, Theorem (1.4.6) and Corollary (1.4.7)]. □

2.1.3 Laplace equation

In this subsection the space V is defined which will be used for the variational formulation of the Cohen model. The idea of the variational formulation is introduced for the Laplace equation, as this is the building block for the time-dependent problem. The theory can be found for example in [OD96, Section 6.6]. For a more rigorous formulation using trace theorems, see [DL88, Section VII.2.2.1].

The Laplace equation, as a PDE on a bounded domain is:

Problem 2.5 (Laplace equation as PDE).
Given $f \in L_2(\Omega)$ and $g \in L_2(\Gamma_N)$, find $u \in H^2(\Omega)$ such that

$$-\Delta u = f \quad \text{in } \Omega, \tag{2.8a}$$
$$u = 0 \quad \text{on } \Gamma_D, \tag{2.8b}$$
$$(\nabla u)^T \boldsymbol{n} = g \quad \text{on } \Gamma_N. \tag{2.8c}$$

In order to give a variational formulation for this problem, first the space V is defined and some of its properties are stated: Note that a rigorous definition of the space V in (2.9) and also of the Neumann boundary condition (2.8c) needs a trace theorem, see for example [DL88, IV.4.3].

Lemma 2.6 (Hilbert space V).
Define the space V by

$$V := \{v \in H^1(\Omega) \,:\, v = 0 \text{ on } \Gamma_D\}, \tag{2.9}$$

and set for any $v, w \in V$:

$$\begin{aligned} (v, w)_V &:= (\nabla v, \nabla w), \\ \|v\|_V &:= \sqrt{(v, v)_V}. \end{aligned} \tag{2.10}$$

Then $(V, (\cdot,\cdot)_V)$ is a Hilbert space. Moreover, on V the V-norm is equivalent to the H^1-norm: there is a constant $C_V > 0$ such that

$$C_V^{-1} \|v\|_{H^1(\Omega)} \le \|v\|_V \le \|v\|_{H^1(\Omega)} \qquad \forall\, v \in V. \tag{2.11}$$

Proof. We just give an outline of the proof: By [DL88, Section VII.2.2.1], the space V is a closed (w.r.t $\|\cdot\|_{H^1(\Omega)}$) subspace of $H^1(\Omega)$. Hence V together with $\|\cdot\|_{H^1(\Omega)}$ is a Banach space. Furthermore, it is easily seen that $(\cdot,\cdot)_V$ is a symmetric, positive-semidefinite bilinear form and hence that $\|\cdot\|_V$ is a seminorm. (Actually $\|\cdot\|_V$ is the so called H^1-seminorm on the space $H^1(\Omega)$.) Thus if (2.11) can be shown, then we would firstly get that $\|\cdot\|_V$ is a norm and that $(\cdot,\cdot)_V$ is positive definite and thus an inner product on V. The completeness of V with respect to $\|\cdot\|_V$ would also follow from the completeness of V with respect to $\|\cdot\|_{H^1(\Omega)}$ proving that V is a Hilbert space. So the only thing left to prove is (2.11).

Now, the right half of the equivalence (2.11) is an immediate consequence of the definition of the H^1-norm, but the left half is the difficult part: This is Poincaré's inequality, which has an easy proof if $\partial\Omega = \Gamma_D$, see [DL88, IV.7.1, Proposition 1]. For the general case where $\Gamma_D \neq \partial\Omega$, the proof makes use of the fact that $H^1(\Omega)$ is *compactly* embedded in $L_2(\Omega)$, see [DL88, IV.7.2, Proposition 2 and Remark 4]. □

If $\mathcal{ER}^1$ holds, and hence the Dirichlet boundary consists of the whole boundary, then the space V is also denoted as $H_0^1(\Omega)$, see for example [Eva98].

We now come to the connection between the two spaces V and $L_2(\Omega)$, they form part of a so-called *Gelfand triple*. For more background on this subject, including the definitions of the dual space, the duality pairing and Gelfand triples, see Section A.4 in the Appendix.

Lemma 2.7 (Gelfand triple).
V is continuously and densely embedded in $L_2(\Omega)$, i.e.

$$\|v\|_0 \leq C_V \|v\|_V \quad \forall v \in V, \tag{2.12}$$

where $C_V > 0$ is the constant from (2.11). Furthermore, $V, L_2(\Omega), V'$ form a Gelfand triple

$$V \hookrightarrow L_2(\Omega) \hookrightarrow V'.$$

Let $\langle \cdot, \cdot \rangle : V' \times V \to \mathbb{R}$ denote the duality pairing of V with V'. Then

$$\langle w, v \rangle = (w, v)_0 \qquad \forall\, w \in L_2(\Omega), v \in V, \tag{2.13}$$

and

$$\|v\|_{V'} \leq C_V \|v\|_0 \quad \forall v \in L_2(\Omega). \tag{2.14}$$

Proof. Again we give just an outline of the proof: By the definition of $H^1(\Omega)$ we have that $H^1(\Omega) \hookrightarrow L_2(\Omega)$ and by (2.11) we get $V \hookrightarrow H^1(\Omega)$. The superposition of these two continuous embeddings gives the continuous embedding $V \hookrightarrow L_2(\Omega)$. By the definition of the H^1-norm we have $\|v\|_0 \leq \|v\|_{H^1(\Omega)}$ for $v \in V$, and together with (2.11) this implies (2.12).

The embedding of V in $L_2(\Omega)$ is also dense, since the embedding of $H_0^1(\Omega)$ in $L_2(\Omega)$ is. Thus Proposition A.26 can be used to obtain the Gelfand triple and (2.13).

It remains to confirm the constant in (2.14): Let $v \in L_2(\Omega)$. Then

$$\begin{aligned} \|v\|_{V'} &= \sup_{0 \neq w \in V} \frac{|\langle v, w \rangle|}{\|w\|_V} = \sup_{0 \neq w \in V} \frac{|(v, w)_0|}{\|w\|_V} \\ &\leq \sup_{0 \neq w \in V} \frac{\|v\|_0 \|w\|_0}{\|w\|_V} \leq \sup_{0 \neq w \in V} \frac{\|v\|_0 C_V \|w\|_V}{\|w\|_V} = C_V \|v\|_0. \end{aligned}$$

□

Now, having defined the space V we can reformulate Problem 2.5 in a variational form. Green's formula is the tool that is used for this.

Lemma 2.8 (Green's formula).
For sufficiently regular functions u, v, we have

$$\int_\Omega (-\Delta u) v = \int_\Omega (\nabla u)^T \nabla v - \int_{\partial\Omega} (\nabla u)^T \boldsymbol{n} v.$$

Supposing that the solution u to Problem 2.5 and an arbitrary function $v \in V$ are sufficiently smooth, we can multiply (2.8a) by v and integrate over the domain, then apply Green's formula, and finally use the Neumann boundary condition (2.8c) to arrive at (2.15) given below. This new formulation makes sense not only for these sufficiently smooth functions but even for functions which are less regular than in Problem 2.5. This leads to the following variational formulation:

Problem 2.9 (Variational form of the Laplace equation).
Given $f \in L_2(\Omega)$ and $g \in L_2(\Gamma_N)$, find $u \in V$ such that

$$(u, v)_V = \langle L, v \rangle \quad \forall v \in V, \tag{2.15}$$

where $L \in V'$ is defined as

$$\langle L, v \rangle := (f, v)_0 + (g, v)_{\Gamma_N} \quad \text{for } v \in V.$$

The standard result of existence and uniqueness for Problem 2.9 is:

Theorem 2.10 (Unique solution of Laplace equation).
If $L \in V'$ then (2.15) has a unique solution $u \in V$ and

$$\|u\|_V = \|L\|_{V'}.$$

Furthermore, if $\mathcal{ER}^1$ holds and if $L \in L_2(\Omega)$ then $u \in H^2(\Omega)$ and

$$\|u\|_{H^2(\Omega)} \leq C\|L\|_0.$$

Proof. The first part follows immediately from the Riesz representation theorem A.24. For the second part see [EG04, Theorem 3.12] for polygonal domains and see [RR04, Theorem 9.53] for C^∞-domains. □

For getting the higher regularity $u \in H^2(\Omega)$ in Theorem 2.10, it is necessary to make assumptions similar to $\mathcal{ER}^1$:

- If in two dimensions the domain is polygonal but not convex, then there is a so called *re-entrant corner*, i.e. a corner with interior angle $> \pi$. In that case the solution u has a singularity at this corner and $u \notin H^2(\Omega)$, see [EG04, Remark 3.13(ii)].
- If the type of boundary condition changes on connected parts of the boundary, then the solution might also not be in $H^2(\Omega)$ anymore, for an example, see [EG04, Remark 3.11(ii)].

2.1.4 Inverse Laplacian

Another concept sometimes used for systems of partial differential equations is the inverse Laplacian, see for example [BGN03].

For $v \in V'$ the inverse Laplacian $\mathcal{G}v$ is defined as the unique solution to the variational problem

$$(\mathcal{G}v, w)_V = \langle v, w\rangle \qquad \forall w \in V. \tag{2.16}$$

Note that $\mathcal{G}$ is linear and has the following properties:

Lemma 2.11 (Inverse Laplacian).
$\mathcal{G} : V' \to V$ is an isometric isomorphism, in particular:

$$\|\mathcal{G}v\|_V = \|v\|_{V'} \qquad \forall v \in V'. \tag{2.17}$$

Furthermore, if $\mathcal{ER}^1$ holds then the inverse Laplacian has a strong stability property: There is $C > 0$ such that

$$\|\mathcal{G}v\|_{H^2(\Omega)} \leq C\|v\|_0 \quad \forall\, v \in L_2(\Omega). \tag{2.18}$$

Proof. Use the Riesz representation theorem A.24 and Theorem 2.10. □

2.2 Time dependent problems

The model problem that is examined in this thesis is time-dependent, hence in this section some standard spaces and tools from functional analysis to deal with time dependent problems are stated.

The time t will always be taken from a bounded interval,

$$t \in [0, T],$$

for some time end point $T < \infty$. The time derivatives are denoted using dots. We write $\dot{v}$, $\ddot{v}$, $\dddot{v}$ for the first, second and third derivative of v with respect to time t.

2.2.1 Spaces for time dependent problems

This subsection gives an introduction to the function spaces that are needed to deal with time-dependent problems. Almost always, the time derivatives that appear in physical equations have a different form than the space derivatives. Therefore, the parameter time cannot be treated as just an extra space dimension. The function spaces usually need different regularity with respect to space than with respect to time.

Let $\mathcal{J} \subseteq \mathbb{R}$ be a time interval. We want to consider functions that depend on both time and space, i.e. functions $v : \mathcal{J} \times \Omega \to \mathbb{R}$. In order to define function spaces for these functions, we reuse the definitions of function spaces from Section 2.1.2. Let X be a Banach space of space-dependent functions $\Omega \to \mathbb{R}$, for example one of those of Section 2.1.2. We identify the functions $v : \mathcal{J} \to X$ with the functions $v : \mathcal{J} \times \Omega \to \mathbb{R}$ by writing

$$v(t)(\boldsymbol{x}) = v(t, \boldsymbol{x}).$$

Let $a < b$. We define the function spaces either on the closed time interval $t \in [a, b]$ or on the open counterpart $t \in (a, b)$:

Lemma 2.12 (The space $C^m([a,b];X)$).
Let $(X, \|\cdot\|_X)$ be a Banach space, $m \in \mathbb{N}$ and $[a, b]$ a bounded interval. Then $C^m([a,b];X)$ is defined as the space of m times continuously differentiable functions from $[a, b]$ into X. This space together with the norm

$$\|v\|_{C^m([a,b];X)} := \sum_{i=0}^{m} \sup_{t\in[a,b]} \|\tfrac{\partial^i}{\partial t^i} v(t)\|_X \tag{2.19}$$

is a Banach space.

Proof. For the definition and statement of the properties of $C^m([a,b];X)$, see [BG87a, 2.10B]. □

For $C^0([a,b];X)$ we also write $C([a,b];X)$ and in this case the definition of the norm simplifies to:

$$\|v\|_{C([a,b];X)} := \max_{t\in[a,b]} \|v(t)\|_X. \tag{2.20}$$

Later we will frequently need the next equality for functions in $C^1([0,T];L_2(\Omega))$:

Lemma 2.13 (An equality for $(\dot{v}, v)_0$).
For any $v \in C^1([0,T];L_2(\Omega))$:

$$(\dot{v}, v)_0 = \frac{1}{2}\frac{d}{dt}\|v\|_0^2. \tag{2.21}$$

Proof. This follows from the definition of the L_2 inner product, since

$$\frac{d}{dt}\|v\|_0^2 = \frac{d}{dt}\int_\Omega v^2 = \int_\Omega 2\dot{v}v = 2(\dot{v}, v)_0.$$ □

The spaces $W_p^m(a,b;X)$

The time dependent version of the Sobolev spaces are:

Lemma 2.14 (The space $W_p^m(a,b;X)$).
Let $m \in \mathbb{N}$ and $1 \le p \le \infty$. Then $W_p^m(a,b;X)$ is defined as the space of functions $(a,b) \to X$ such that their X-norm lies in $W_p^m((a,b))$. For $v \in W_p^m(a,b;X)$ set

$$w : \left\{ \begin{array}{rcl} (a,b) & \to & \mathbb{R} \\ t & \mapsto & \|v(t)\|_X, \end{array} \right.$$

and use this to define the norm

$$\|v\|_{W_p^m(a,b;X)} := \|w\|_{W_p^m((a,b))}.$$

When equipped with this norm, $W_p^m(a,b;X)$ is a Banach space.

Proof. See [BG87a, 2.10D]. □

As before for the space dependent functions, the special cases $m = 0$ and $p = 2$ are also denoted differently:

$$L_p(a,b;X) := W_p^0(a,b;X), \qquad H^m(a,b;X) := W_2^m(a,b;X).$$

The norms for these spaces are also defined in this way.

Lemma 2.15 (Embeddings of time-dependent function spaces).
i) Let X, Y be Banach spaces such that $X \hookrightarrow Y$. Then for any $m \in \mathbb{N}$ and $1 \le p \le \infty$,

$$W_p^m(a,b;X) \hookrightarrow W_p^m(a,b;Y),$$
$$C^m([a,b];X) \hookrightarrow C^m([a,b];Y).$$

ii) Let X be a Banach space. If $l, m \in \mathbb{N}$ and $1 \le p, q \le \infty$ are such that $W_p^l((a,b)) \hookrightarrow W_q^m((a,b))$ then

$$W_p^l(a,b;X) \hookrightarrow W_q^m(a,b;X).$$

Proof. This follows from the definitions. See also [BG87a, 2.10B, 2.10D]. □

As an example of the use of this lemma, observe that (since the time-interval $(0,T)$ is one-dimensional), the Sobolev embedding Theorem 2.4 implies that $H^2((a,b)) \hookrightarrow W_\infty^1((a,b))$, and hence

$$H^2(0,T;L_2(\Omega)) \hookrightarrow W_\infty^1(0,T;L_2(\Omega)), \tag{2.22}$$

an embedding that will come to use later in Chapter 6.

The space $\mathcal{W}(0,T;V,V')$

There is another important space of time dependent functions, which we will define now. It is the space in which an important class of time-dependent problems has a unique solution, see later in Theorem 2.19. The definition is taken from [DL88, XVIII.1.2 Definition 4] and [Wlo87, Definition 25.3].

Definition 2.16 (The space $\mathcal{W}(0,T;V,V')$).
Let $V \hookrightarrow L_2(\Omega) \hookrightarrow V'$ be the Gelfand triple from Lemma 2.7. Define

$$\mathcal{W}(0,T;V,V') := \{u : u \in L_2(0,T;V), \dot{u} \in L_2(0,T;V')\}.$$

Associated with this space is the norm given by

$$\|u\|_{\mathcal{W}(0,T;V,V')} := \left(\|u\|^2_{L_2(0,T;V)} + \|\dot{u}\|^2_{L_2(0,T;V')}\right)^{1/2}.$$

Lemma 2.17 (Properties of $\mathcal{W}(0,T;V,V')$).
i) $\mathcal{W}(0,T;V,V')$ is a Hilbert space.
ii) If $w \in \mathcal{W}(0,T;V,V')$, then w has a function in its equivalence class, which is continuous in time and

$$\mathcal{W}(0,T;V,V') \hookrightarrow C([0,T];L_2(\Omega)). \tag{2.23}$$

iii) Let $w \in \mathcal{W}(0,T;V,V')$. Then

$$\langle \dot{w}, w \rangle = \frac{1}{2}\frac{d}{dt}\|w\|_0^2. \tag{2.24}$$

iv) Let $w \in \mathcal{W}(0,T;V,V')$. Then

$$\langle \dot{w}, \mathcal{G}w \rangle = \frac{1}{2}\frac{d}{dt}\|w\|_{V'}^2, \tag{2.25}$$

where $\mathcal{G}$ is the inverse Laplacian, defined in Subsection 2.1.4.

Proof. i) see [DL88, XVIII.1.2 Proposition 6] or [Wlo87, Theorem 25.4].
ii) See [Wlo87, Theorem 25.5]. Similar statements can be found in [DL88, XVIII.1.2 Theorem 1 and Remark 4] and [Eva98, 5.9 Theorem 3].
iii) See [Tem97, Chapter II, Lemma 3.2].
iv) First we get from (2.16) and (2.17), for any $w \in \mathcal{W}(0,T;V,V')$,

$$(w, \mathcal{G}w)_0 = (\mathcal{G}w, \mathcal{G}w)_V = \|\mathcal{G}w\|_V^2 = \|w\|_{V'}^2.$$

To continue, for any $w \in L_2(0,T;V)$ with $\dot{w} \in L_2(0,T;V)$, we have that

$$\frac{d}{dt}(w, \mathcal{G}w)_0 = (\dot{w}, \mathcal{G}w)_0 + (w, \mathcal{G}\dot{w})_0 = 2(\dot{w}, \mathcal{G}w)_0 = 2\langle \dot{w}, \mathcal{G}w \rangle,$$

since $(w, \mathcal{G}\dot{w})_0 = (\mathcal{G}w, \mathcal{G}\dot{w})_V = (\mathcal{G}\dot{w}, \mathcal{G}w)_V = (\dot{w}, \mathcal{G}w)_0$. By combining the above equations, we get (2.25), if $\dot{w} \in L_2(0,T;V)$. The result extends by density to any $w \in \mathcal{W}(0,T;V,V')$. □

2.2.2 Unique solution of ODEs

In Chapter 4 we will need an an existence and uniqueness result for ordinary differential equations. As we do not need the full generality of the Picard-Lindelöf theorem, we quote a result for the special case of *linear* ODEs:

Proposition 2.18 (Unique solution for ODEs).
Let Y be a Banach space, $A \in C([0,T];\mathscr{L}(Y,Y))$, $b \in C([0,T];Y)$ and $y_0 \in Y$. Then the linear inhomogeneous initial value problem

$$\begin{aligned} \dot{y} &= A(t)y + b(t), \\ y(0) &= y_0, \end{aligned}$$

has a unique solution $y \in C^1([0,T];Y)$.

A proof can be found for example in [Ama90, Theorem (7.9) and Remark (7.10) c)].

2.2.3 Linear parabolic equations

Another result of existence and uniqueness that will be needed in Chapter 4 is the following theorem. For the proof see for example [Wlo87, §26] or [RR04, Theorem 11.3].

Theorem 2.19 (Existence and uniqueness for linear parabolic problem).
Suppose that $L \in L_2(0,T;V')$ and $\breve{u} \in L_2(\Omega)$. Then the problem

$$\langle \dot{u}, v \rangle + (u, v)_V = \langle L, v \rangle \quad \forall v \in V \ \forall t \in (0,T) \ \textit{a.e.} \tag{2.26a}$$

$$u(0) = \breve{u} \tag{2.26b}$$

has a unique solution $u \in \mathcal{W}(0,T;V,V')$.

Remark 2.20.
Note that the initial condition (2.26b) has sense for $u \in \mathcal{W}(0,T;V,V')$, i.e. $u(0)$ is well-defined, since there is a continuous function in the equivalence class of u:

$$u \in C([0,T];L_2(\Omega)),$$

by (2.23).

2.3 Space discretisation

In this section we present some elements from standard finite element theory and again define the notation used in the later chapters.

2.3.1 Mesh and finite element spaces

In Definition 2.1 Ω was required to have either a smooth or a polyhedral boundary. However for the space discretisation used in this thesis the domain is always assumed to be polyhedral. This is done for simplicity.

Assumption 2.21.
For any $h \in (0,\hat{h}]$, construct a mesh $\mathcal{E}_h$,which is a finite set of elements that satisfies the following conditions:

- Each finite element is a non-empty open set in $\mathbb{R}^d$ and $E \in \mathcal{E}_h$ is a *simplex*, i.e. an interval if $d=1$, a triangle if $d=2$ or a tetrahedron if $d=3$.
- The elements form a partition of the domain Ω, i.e. for any h

$$\bigcup_{E\in\mathcal{E}_h} \bar{E} = \bar{\Omega},$$
$$E_1 \cap E_2 = \varnothing \quad \forall E_1, E_2 \in \mathcal{E}_h \text{ with } E_1 \neq E_2.$$

- The meshes are geometrically conformal, there are no "hanging nodes": If for elements $E_1, E_2 \in \mathcal{E}_h$ the set $\bar{E}_1 \cap \bar{E}_2$ has positive $(d-1)$-dimensional measure, then it is an edge $(d=2)$ or a face $(d=3)$ of both E_1 and E_2. (Compare [EG04, Definition 1.55, Remark 1.56].)
- The type of boundary condition does not change on an edge or face: If f is an edge $(d=2)$ or a face $(d=3)$ of an element, then $f \subseteq \partial\Omega$ implies either $f \subseteq \Gamma_D$ or $f \subseteq \overline{\Gamma}_N$.
- The element diameters are bounded by h, i.e.

$$\forall h\, \forall E \in \mathcal{E}_h \quad h_E \leq h,$$

  where h_E is the *diameter* of element E:

$$h_E := \sup_{\boldsymbol{x},\boldsymbol{y}\in E} |\boldsymbol{x}-\boldsymbol{y}|.$$

- The mesh family is *quasi-uniform*, i.e. there is a constant $\tau > 0$ such that

$$\forall h\, \forall E \in \mathcal{E}_h \quad \rho_E \geq \tau h,$$

  where ρ_E is the diameter of the largest ball that can be inscribed in E.

Definition 2.22 (Finite element spaces).
Let $\mathcal{E}_h$ be a mesh for a certain $h > 0$. Let also $r \geq 1$. The finite element spaces W^h and V^h are defined by

$$\begin{aligned} W^h &:= \{v \in C^0(\bar{\Omega}) \ : \ \forall E \in \mathcal{E}_h \ v|_E \in P_r(E)\}, \\ V^h &:= \{v \in W^h \ : \ v|_{\Gamma_D} = 0\}, \end{aligned} \tag{2.27}$$

where $P_r(E)$ is the set of polynomial functions in d variables of total degree less than or equal to r on $E \subseteq \mathbb{R}^d$.

Definition 2.23 (Lagrange nodes).
For $E \in \mathcal{E}_h$ let $\boldsymbol{a}_{E,i}, i \in \{0, \dots, d\}$ denote the corner points of the simplex E. The set of *local Lagrange nodes* of E is defined as

$$\mathcal{LN}_E^r := \left\{ \sum_{i=0}^{d} \frac{l_i}{r} \boldsymbol{a}_{E,i} \ : \ l_0, l_1, \dots, l_d \in \{0, \dots, r\} \wedge \sum_{i=0}^{d} l_i = r \right\},$$

where each tuple $(l_0, l_1, \dots, l_d)$ with sum equal to r gives rise to one Lagrange node. Note that the tuples $(\frac{l_0}{r}, \frac{l_1}{r}, \dots, \frac{l_d}{r})$ form the *barycentric coordinates* of the Lagrange nodes, for more information, see [EG04, Section 1.2.3]. The set of *global Lagrange nodes* is then defined by

$$\mathcal{LN}^r := \bigcup_{E \in \mathcal{E}_h} \mathcal{LN}_E^r$$

Let $N_W := |\mathcal{LN}^r|$ be the total number of Lagrange nodes and $N_V := |\mathcal{LN}^r \setminus \Gamma_D|$ the number of Lagrange nodes which are not on the Dirichlet boundary. Denote the Lagrange nodes by $\boldsymbol{a}_j$ and enumerate them in such a way that

$$\mathcal{LN}^r = \{\boldsymbol{a}_j \ : \ j = 1, \dots, N_W\}, \tag{2.28}$$

and that for $j \in \{1, \dots, N_W\}$,

$$\boldsymbol{a}_j \in \Gamma_D \iff j > N_V.$$

Lemma 2.24 (Lagrange basis functions).
With the definitions from above:

1. *For any $j = 1, \dots, N_W$, the conditions*

$$\phi_j(\boldsymbol{a}_i) = \delta_{ij} \quad \forall\, i \in \{1, \dots, N_W\} \tag{2.29}$$

   uniquely define a function $\phi_j \in W^h$.

2. *The set $\{\phi_i \ : \ i = 1, \dots, N_W\}$ forms a basis of W^h and $\{\phi_i \ : \ i = 1, \dots, N_V\}$ forms a basis of V^h.*

3. *W^h is $H^1(\Omega)$-conformal and V^h is V-conformal:*

$$W^h \subseteq H^1(\Omega), \qquad V^h \subseteq V.$$

Proof. This is standard finite element theory, see for example [EG04, Proposition 1.34 and Section 1.4.5]. For the conformality, see [EG04, Proposition 1.74]. □

Remark 2.25.
Although the finite element space depends not only on h but also on the polynomial degree r, we do not write it as another index $V^{h,r}$. It is always assumed that r is fixed.

Remark 2.26 (Why Lagrange basis functions?).
The specific choice of Lagrange basis functions here is made purely for practical reasons: the ALBERTA library, which was chosen for the numerical experiments, uses the Lagrange basis functions up to the polynomial degree of $r = 4$, see later in Section 7.2.

2.3.2 Inverse estimates

Since in the finite dimensional space V^h all norms are equivalent, any norm on V^h can be bounded by any other norm. However the equivalence constants can depend on h. In this subsection so-called inverse estimates are stated, which give the form of h-dependence in some cases. The bounds are mostly in the opposite direction, when compared to the Sobolev embedding, Theorem 2.4.

Theorem 2.27 (Inverse Estimates).
Suppose that $1 \leq p \leq \infty$, $1 \leq q \leq \infty$ *and* $0 \leq m \leq l$, *such that* $V^h \subseteq W_p^l(\Omega) \cap W_q^m(\Omega)$. *Then there is a constant* $C > 0$, *independent of* h *such that*

$$\|v\|_{W_p^l(\Omega)} \leq C\, h^{m-l+\min(0,\frac{d}{p}-\frac{d}{q})} \|v\|_{W_q^m(\Omega)},$$

for all $v \in V^h$.

Proof. This is a special case of [Cia02, Theorem 3.2.6] (or [BS94, Theorem (4.5.11)], which has the technical restriction of $h < 1$). □

As the finite element space V^h contains functions which are in general not continuously differentiable across element boundaries, the restriction $V^h \subseteq W_p^l(\Omega) \cap W_q^m(\Omega)$ is quite severe. Therefore, the inverse estimates are usually stated in a more general form, where the W_p^l-norm and W_q^m-norm is applied on each element and not on the whole domain. For the sake of simplicity and because we do not need the general form, the inverse estimates are stated the way they are.

For later, we note some special cases of the inverse estimates:

$$\|v\|_{L_\infty(\Omega)} \leq Ch^{-d/2} \|v\|_0 \quad \forall\, v \in V^h.$$

$$\|v\|_V \leq Ch^{-1} \|v\|_0 \quad \forall\, v \in V^h. \tag{2.30}$$

2.3.3 Interpolation

Nodal interpolation is used for one of the five discrete schemes that are considered in Chapter 5.

Definition 2.28 (Nodal interpolation).
The nodal interpolation operator $I^h : C^0(\bar{\Omega}) \to W^h$ is defined by

$$I^h v := \sum_{j=1}^{N_W} v(\boldsymbol{a}_j)\phi_j, \tag{2.31}$$

for any $v \in C^0(\bar{\Omega})$.

Note that as a consequence of this definition, nodal interpolation preserves homogeneous Dirichlet boundary conditions: $I^h(C^0(\bar{\Omega}) \cap V) \subseteq V^h$.

We state a standard result of the approximation error of nodal interpolation, i.e. of how close the interpolant is to the original function:

Lemma 2.29 (Approximation error of interpolation).
Suppose that $\frac{d}{2} - 1 < l \leq r$. *Then for any* $v \in H^{l+1}(\Omega) \cap V$:

$$\|v - I^h v\|_0 \leq Ch^{l+1} \|v\|_{H^{l+1}(\Omega)},$$
$$\|v - I^h v\|_V \leq Ch^{l} \|v\|_{H^{l+1}(\Omega)}.$$

Proof. This is a special case of [EG04, Corollary 1.109, Example 1.106]. Observe that l has to be chosen large enough so that $H^{l+1}(\Omega)$ is continuously embedded in $C^0(\bar{\Omega})$. □

One problem with the nodal interpolation operator is its lack of stability with respect to the L_2-norm. To demonstrate this, we give an example.

Example 2.30 (About stability of the interpolation operator).
We are going to show the following:
There is no $C > 0$, independent of h, such that

$$\|I^h v\|_0 \leq C\|v\|_0 \quad \forall v \in C^\infty[-1,1].$$

Note that we choose the space of infinitely differentiable functions here. Since there is no stability in this space, as we will see, we cannot have stability in any bigger, less regular space containing $C^\infty[-1,1]$.

To give the proof, let $\Omega = (-1,1)$ in 1D be discretised with a mesh with equidistant nodes: $\ldots, -2h, -h, 0, h, 2h, \ldots$ using piecewise polynomials, $r = 1$. For each $\epsilon \in (0,h)$ choose a function $v_\epsilon \in C^\infty[-1,1]$ such that $0 \leq v_\epsilon(x) \leq 1$, $v_\epsilon(0) = 1$ and $v_\epsilon(x) = 0$ for $|x| \geq \epsilon$. Such a function exists, a possible choice would be, for example,

$$v_\epsilon(x) := \begin{cases} \exp(1 - 1/(1-(x/\epsilon)^2)) & \text{if } |x| < \epsilon, \\ 0 & \text{otherwise.} \end{cases}$$

See Figure 2.1 for a sketch of v_ϵ and its interpolant $I^h v_\epsilon$. Now since $\epsilon < h$ we have that

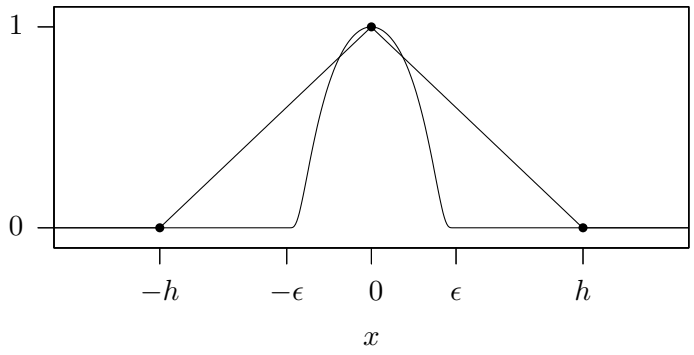

Figure 2.1: Function v_ϵ and its interpolant $I^h v_\epsilon$ from Example 2.30.

$$\|I^h v_\epsilon\|_0 = \left(2\int_0^h \left(1 - \frac{x}{h}\right)^2 dx\right)^{1/2} = \sqrt{2h/3},$$

but on the other hand $\|v_\epsilon\|_0 \leq \sqrt{2\epsilon}$. Thus for each $C > 0$ there is an $\epsilon > 0$ such $\|I^h v_\epsilon\|_0 > C\|v_\epsilon\|_0$, which is what we wanted to prove.
An equivalent statement to the above is: There is no $C > 0$ such that

$$\|v - I^h v\|_0 \leq C\|v\|_0 \quad \forall v \in C^\infty[-1,1].$$

Scott-Zhang interpolation

Because of the main drawback of the nodal interpolation I^h, i.e. that it can only be applied to continuous functions, other types of interpolation operators have been defined and studied. Two such examples are the Clément operator and Scott-Zhang interpolation, see [EG04, Section 1.6].

Here, we quote properties of the Scott-Zhang interpolation, without giving a definition of it. This type of interpolation will only be needed in this chapter; in the next subsection we will use it to prove some properties of the L_2- and elliptic projections.

Lemma 2.31 (Scott-Zhang interpolation).
There is a projection operator $\mathcal{SZ}^h : H^1(\Omega) \to W^h$ which preserves homogeneous boundary conditions such that $\mathcal{SZ}^h(V) \subseteq V^h$. This interpolation operator is stable in $H^1(\Omega)$,

$$\|\mathcal{SZ}^h v\|_{H^1(\Omega)} \leq C\|v\|_{H^1(\Omega)}, \tag{2.32}$$

and its approximation properties are, for $l \in \{0, \dots, r\}$,

$$\|v - \mathcal{SZ}^h v\|_0 \leq Ch^{l+1} \|v\|_{H^{l+1}(\Omega)} \quad \forall\, v \in H^{l+1}(\Omega), \tag{2.33}$$

$$\|v - \mathcal{SZ}^h v\|_{H^1(\Omega)} \leq Ch^{l} \|v\|_{H^{l+1}(\Omega)} \quad \forall\, v \in H^{l+1}(\Omega). \tag{2.34}$$

Proof. This is a specialisation of [EG04, Lemma 1.130]. □

2.3.4 L_2-projection and elliptic projection

This subsection defines two projection operators that map into V^h, namely the L_2-projection and the elliptic projection. Both are orthogonal projections from a bigger space onto the subspace V^h. Some stability and approximation properties of these operators are included.

The L_2-projection will be used for calculating the discrete initial solution, see later in Section 5.2. The elliptic projection will be needed for proving the error estimate in Chapter 6.

Definition 2.32 (L_2- and elliptic projection).
For $v \in V$ define the *elliptic projection* $\Pi_V^h v \in V^h$ via

$$(\Pi_V^h v, w)_V = (v, w)_V \qquad \forall w \in V^h.$$

Similarly, for $v \in L_2(\Omega)$ define the *L_2-projection* $\Pi_0^h v \in V^h$ via

$$(\Pi_0^h v, w)_0 = (v, w)_0 \qquad \forall w \in V^h.$$

Note that the form of the elliptic projection is usually problem dependent and comes from the form of the space derivatives of the problem under study. Thus the elliptic projection is sometimes defined using the full H^1 inner product $(\cdot,\cdot)_V + (\cdot,\cdot)_0$ instead of only the inner product $(\cdot,\cdot)_V$ of V, see for example [EG04]. The next lemma is [EG04, Lemma 1.113], modified only to have the V-norm instead of the (equivalent) H^1-norm:

Lemma 2.33 (Stability of L_2- and elliptic projections).
There is $C > 0$, independent of h, such that

$$\begin{aligned}
\|\Pi_0^h v\|_0 &\leq \|v\|_0 && \forall v \in L_2(\Omega),\\
\|\Pi_V^h v\|_V &\leq \|v\|_V && \forall v \in V,\\
\|\Pi_0^h v\|_V &\leq C\|v\|_V && \forall v \in V.
\end{aligned}$$

Proof. Using the definition of the L_2-projection and the Cauchy-Schwarz inequality, we get

$$\|\Pi_0^h v\|_0^2 = (\Pi_0^h v, \Pi_0^h v)_0 = (v, \Pi_0^h v)_0 \leq \|v\|_0 \cdot \|\Pi_0^h v\|_0.$$

Dividing by $\|\Pi_0^h v\|_0$ yields the first inequality.

Analogously, for the elliptic projection,

$$\|\Pi_V^h v\|_V^2 = (\Pi_V^h v, \Pi_V^h v)_V = (v, \Pi_V^h v)_V \leq \|v\|_V \cdot \|\Pi_V^h v\|_V,$$

which gives the second inequality.

The proof of the third inequality follows [EG04, Exercise 1.17], for another proof, see [Bra97, Corollary 7.8]. Let $v \in V$. For any $z \in V^h$ we have by the triangle inequality

$$\|\Pi_0^h v\|_V \leq \|\Pi_0^h v - z\|_V + \|z\|_V. \tag{2.35}$$

The choice of z depending on v will be made later. To obtain a bound for the term $\|\Pi_0^h v - z\|_V$, we first find a bound for $\|\Pi_0^h v - z\|_0$:

$$\begin{aligned}\|\Pi_0^h v - z\|_0^2 &= (\Pi_0^h v - z, \Pi_0^h v - z)_0 = (\Pi_0^h v, \Pi_0^h v - z)_0 - (z, \Pi_0^h v - z)_0 \\ &= (v, \Pi_0^h v - z)_0 - (z, \Pi_0^h v - z)_0 = (v - z, \Pi_0^h v - z)_0 \\ &\leq \|v - z\|_0 \cdot \|\Pi_0^h v - z\|_0,\end{aligned}$$

whence

$$\|\Pi_0^h v - z\|_0 \leq \|v - z\|_0.$$

Since $\Pi_0^h v - z \in V^h$ we can use the inverse estimate (2.30), which gives

$$\|\Pi_0^h v - z\|_V \leq Ch^{-1}\|\Pi_0^h v - z\|_0.$$

Combining all these with (2.35) gives

$$\|\Pi_0^h v\|_V \leq Ch^{-1}\|v - z\|_0 + \|z\|_V. \tag{2.36}$$

The remaining problem is to choose a function $z \in V^h$ for which these terms can be bounded by $\|v\|_V$. Choosing the nodal interpolant $z = I^h v$ would require that v is continuous in order to apply Lemma 2.29 and this is the case only for $d = 1$. Therefore this choice for z would not be useful in general. The Scott-Zhang interpolant $\mathcal{SZ}^h v$, however, has the necessary properties: The stability

$$\|\mathcal{SZ}^h v\|_V \leq C\|v\|_V,$$

follows from (2.32) since $v, \mathcal{SZ}^h v \in V$ and because of the equivalence of V-norm and H^1-norm (2.11). The approximation property is a consequence of (2.33),

$$\|v - \mathcal{SZ}^h v\|_0 \leq Ch\|v\|_V.$$

Thus, choosing $z = \mathcal{SZ}^h v$ and using stability and approximation in (2.36), we get

$$\|\Pi_0^h v\|_V \leq C\|v\|_V,$$

which is what we wanted to prove. □

Lemma 2.34 (Approximation error for L_2- and elliptic projections). *Suppose that $\mathcal{ER}^q$ holds and that $0 \leq l \leq r$. There is $C > 0$, independent of h, such that*

$$\|v - \Pi_0^h v\|_0 \leq Ch^{l+1}\|v\|_{H^{l+1}(\Omega)} \qquad \forall v \in H^{l+1}(\Omega), \tag{2.37}$$

$$\|v - \Pi_V^h v\|_0 \leq Ch^{l+q}\|v\|_{H^{l+1}(\Omega)} \qquad \forall v \in V \cap H^{l+1}(\Omega), \tag{2.38}$$

$$\|v - \Pi_V^h v\|_V \leq Ch^{l}\|v\|_{H^{l+1}(\Omega)} \qquad \forall v \in V \cap H^{l+1}(\Omega). \tag{2.39}$$

Note that the 'missing' approximation property

$$\|v - \Pi_0^h v\|_V \leq Ch^{l}\|v\|_{H^{l+1}(\Omega)} \quad \forall v \in V \cap H^{l+1}(\Omega),$$

holds as well, see [EG04, Proposition 1.134], but it is not needed for this thesis.

Proof of Lemma 2.34. The proof is very similar to the one of [EG04, Proposition 1.134], it only needs to be adapted for the different definition of the elliptic projection and the use of the V-norm instead of the H^1-norm. We start by proving (2.39). First observe that the definition of the elliptic projection directly implies this orthogonality equation:

$$(v - \Pi_V^h v, w)_V = 0 \quad \forall\, w \in V^h. \tag{2.40}$$

Using this we get the following chain of inequalities: For any $v^h \in V^h$,

$$\begin{aligned}\|v - \Pi_V^h v\|_V^2 &= (v - \Pi_V^h v, v - \Pi_V^h v)_V = (v - \Pi_V^h v, v - v^h)_V \\ &\leq \|v - \Pi_V^h v\|_V \cdot \|v - v^h\|_V.\end{aligned}$$

Thus

$$\|v - \Pi_V^h v\|_V \leq \|v - v^h\|_V \qquad \forall\, v^h \in V^h.$$

Choosing $v^h := \mathcal{SZ}^h v$ and using (2.34) yields (2.39).

The proof of (2.37) is directly analogous, using $(\cdot,\cdot)_0$ and $\|\cdot\|_0$ instead of $(\cdot,\cdot)_V$ and $\|\cdot\|_V$ and finally using (2.33).

Now to the proof of (2.38): For $q = 0$, this is a direct consequence of (2.39) and (2.12). Therefore suppose now that $q = 1$. We use the technique of the Aubin-Nitsche Lemma for the proof (see for example [EG04, Section 2.3.4]). Let $v \in V$ and set $e = v - \Pi_V^h v$. Define $z \in V$ as the solution to the dual problem

$$(z, w)_V = (e, w)_0 \quad \forall\, w \in V. \tag{2.41}$$

Then by Theorem 2.10 and using the elliptic regularity $\mathcal{ER}^1$, we have that $z \in H^2(\Omega)$ and

$$\|z\|_{H^2(\Omega)} \leq C\|e\|_0. \tag{2.42}$$

Now

$$\begin{aligned}\|e\|_0^2 = (e,e)_0 &\overset{(2.41)}{=} (z,e)_V \overset{(2.40)}{=} (z - \Pi_V^h z, e)_V \leq \|z - \Pi_V^h z\|_V \cdot \|e\|_V \\ &\overset{(2.39)}{\leq} Ch\|z\|_{H^2(\Omega)} \cdot h^l \|v\|_{H^{l+1}(\Omega)} \overset{(2.42)}{\leq} Ch^{l+1}\|e\|_0 \cdot \|v\|_{H^{l+1}(\Omega)},\end{aligned}$$

whence

$$\|e\|_0 \leq Ch^{l+1}\|v\|_{H^{l+1}(\Omega)},$$

which is (2.38). □

Chapter 3

The Cohen Model for Non-Fickian Diffusion

In this chapter we state the PDE-form of the problem for this thesis, see Section 3.1. Then in Sections 3.2 and 3.3 we give notes on the physical interpretation of the model and the available literature.

In Section 3.4 we will see that by rescaling the model some of its parameters can be removed, making the notation shorter.

We conclude the chapter by presenting several choices for a nonlinear term in the model and by proving some properties of this nonlinearity in Section 3.5.

3.1 The Cohen model

The model was proposed by Cohen and White in [CW91]. We state it here in a multidimensional version on a bounded domain, similar to [CWW95]. For simplicity, we refer to the problem as the *Cohen model*, although many more people have contributed to the formulation and understanding, see Section 3.3. Recall from Definition 2.1 and 2.2 that Ω is an open, bounded domain with its boundary split into a Dirichlet part Γ_D and a Neumann part Γ_N.

Problem 3.1 (The Cohen model).
Find $u = u(t, \boldsymbol{x})$, $\sigma = \sigma(t, \boldsymbol{x})$ such that

$$\dot{u} - \nabla^T D \nabla u = f + \nabla^T K \nabla \sigma \quad \text{in } \Omega \times [0,T] \tag{3.1a}$$

$$\dot{\sigma} + \gamma(u)\sigma = \mu\, u + \nu\, \dot{u} + \hat{f} \quad \text{in } \Omega \times [0,T] \tag{3.1b}$$

$$u = u_D \quad \text{on } \Gamma_D \times [0,T] \tag{3.1c}$$

$$(D\nabla u + K \nabla \sigma)^T \boldsymbol{n} = g \quad \text{on } \Gamma_N \times [0,T] \tag{3.1d}$$

$$u = \breve{u} \quad \text{at } t = 0 \tag{3.1e}$$

$$\sigma = \breve{\sigma} \quad \text{at } t = 0 \tag{3.1f}$$

where $D, K, \mu > 0$ are positive constants, ν a real constant and $f = f(t, \boldsymbol{x})$, $\hat{f} = \hat{f}(t, \boldsymbol{x})$, $g = g(t, \boldsymbol{x})$, $u_D = u_D(t, \boldsymbol{x})$, $\breve{u} = \breve{u}(\boldsymbol{x})$, and $\breve{\sigma} = \breve{\sigma}(\boldsymbol{x})$ are given functions.

The function γ in (3.1b) is a nonlinear function, for which we do not give a specific form. Instead we assume throughout this thesis that γ satisfies Assumption 3.4 given below. The original model from [CWW95] has a specific form for γ that involves a tanh-function, see Lemma 3.6.

Remark 3.2.
The terms f and $\hat{f}$ in Problem 3.1 are not part of the physical model, but introduced

here to be able to prescribe exact solutions u, σ to the equations. This is important for testing the implementation and showing convergence of the methods.

Remark 3.3 (The choice of $\nu = 0$).
The term $\nu\dot{u}$ in (3.1b) was included in the model in [CW91] and in [HZ96]. However in both articles, μ and ν are taken to be dependent on u. In [CWW95] however, $\nu = 0$ was chosen. Because it would lead to more theoretical difficulties, we follow the example of [CWW95] and take $\nu = 0$ for this thesis.

Assumption 3.4 (Properties of nonlinear function γ).
Function $\gamma : \mathbb{R} \to \mathbb{R}$ must have these properties:

- γ is twice continuously differentiable: $\gamma \in C^2(\mathbb{R})$.
- There are constants $\check{\gamma}$ and $\hat{\gamma}$ such that

$$0 < \check{\gamma} \leq \gamma(u) \leq \hat{\gamma} \quad \forall u \in \mathbb{R}. \tag{3.2}$$

- There is a constant C'_γ such that

$$0 \leq \gamma'(u) \leq C'_\gamma \quad \forall u \in \mathbb{R}. \tag{3.3}$$

- There is a constant C''_γ such that

$$|\gamma''(u)| \leq C''_\gamma \quad \forall u \in \mathbb{R}. \tag{3.4}$$

In Subsection 3.5.1 we will present three possible choices for γ that satisfy these assumptions.

3.2 Physical interpretation of the model

Problem 3.1 is supposed to model the concentration u of certain substances inside a polymer. If this system could be described by the standard Fick's law, then u would be modelled by the heat equation:

$$\dot{u} - D\Delta u = 0.$$

In order to explain experimental results, see [TW78], this equation was extended in [CW91]. Another term involving a new unknown function σ was added to get (3.1a). Because of similarities to the theory of viscoelastic materials, σ is often called 'stress'. A second equation (3.1b) was added to describe how σ evolves with time. Note that (3.1b) can be solved explicitly for σ. With the simplifying assumptions that $\sigma(0) = 0$ and $\nu = \hat{f} = 0$ we get

$$\sigma = \int_0^t \exp\Big(-\int_s^t \gamma(u(\xi))d\xi\Big)\, \mu u(s)\, ds.$$

Inserting this in (3.1a) gives

$$\dot{u} - \nabla^T D \nabla u = f + \nabla^T K \nabla \int_0^t \exp\Big(-\int_s^t \gamma(u(\xi))d\xi\Big)\, \mu u(s)\, ds.$$

It can be seen that the change in u depends on the whole history of u, i.e. all values $u(s)$ for $s \in (0, t)$. This is a property of viscoelastic materials, they possess a 'memory' of their history.

The function γ in (3.1b) models a phase change of the polymer, depending on the concentration u. If u is below a critical value u_a, then the polymer is in a so-called *glassy* state, which is characterised by a small constant value of γ. If, on the other hand, u is above u_a, then the polymer is in a so-called *rubbery* state and the value of γ is also constant but considerably higher than in the glassy state, see [CW91]. In order to avoid choosing a discontinuous jump function for γ, a small transition region around u_a was introduced and γ was taken to be constant (or almost constant) outside of this transition region. This way γ can still be infinitely differentiable, see Subsection 3.5.1.

3.3 Notes on literature

The model was proposed by Cohen and White in [CW91], first deriving a general form, where the coefficients D, K, μ and ν are allowed to depend on u. This article also introduces the tanh-form of γ, given by Lemma 3.6. Later the special case is studied where D, K, μ are constant and $\nu = 0$.

This special case is also studied in [CWW95] on a two-dimensional bounded domain.

In [Edw95], Edwards obtains solutions with a wave front advancing at constant speed on the domain $\Omega = (0, \infty)$. He replaces the *tanh*-form of γ by a discontinuous function taking only two values. This is generalised in [SE04] giving travelling wave solutions. In that article D and K are dependent on u, μ and ν are constant and there is a more general form for γ.

Rivière and Shaw have modified the problem in [RS06] (see also [RS05] for an extended version) and obtained a priori estimates for a discontinuous Galerkin finite element method applied to their modified problem. The equations they are studying are

$$\begin{aligned} \dot{u} - \nabla^T D \nabla u - \nabla^T K \boldsymbol{\sigma} &= f, \\ \dot{\boldsymbol{\sigma}} + \gamma(u)\boldsymbol{\sigma} &= \mu \nabla u. \end{aligned}$$

That is, the stress $\boldsymbol{\sigma}$ here is vector-valued and the first equation contains the divergence of $\boldsymbol{\sigma}$ instead of the Laplacian of σ as in (3.1a). Moreover the second equation now has the gradient of u on the right-hand side. This modified problem appears to show similar behaviour to Problem 3.1 but it is easier to prove estimates for it, see also the beginning of Chapter 6.

In [Vor08] Vorotnikov applies the technique of 'dissipative solutions' to the problem to obtain a certain type of weak solution in the whole space $\mathbb{R}^2$ or $\mathbb{R}^3$.

Amann shows in [Ama93] local existence and uniqueness for a generalisation of the problem. For example, the factors D, K and μ are allowed to depend on the solution and also the function γ is not assumed to be bounded. This paper applies very general techniques for quasilinear evolution equations to many different problems to show the power of these methods.

A few years before, in [Ama91], Amann had studied the special problem that we are looking at and has shown global existence and uniqueness. This result is the most relevant for the form of equations that are studied in this thesis. Therefore we state it in Appendix B.

Hu and Zhang show in [HZ96] global existence and uniqueness for the generalised problem. They assume, instead of the term μu in the second equation a bounded term $\mu(u)u$ where $\mu(u) = 0$ for any u which is greater than a fixed value. The method used for their proof is the Banach fixed point theorem.

In this thesis we will also look at the question of existence and uniqueness of solutions. Chapter 4 gives a uniqueness result in 1D for a variational formulation of the

problem.

3.4 Rescaling the equations

In order to simplify the problem without losing any of its properties, the variables can be rescaled:

Lemma 3.5 (Rescaling the Cohen model).
By scaling the time variable t, the space variable $\boldsymbol{x}$ and also σ in Problem 3.1, we can choose $D = K = \mu = 1$. Furthermore, suppose that u_a is a special value of u (for example, a Dirichlet boundary value or a certain value of function γ). By also scaling u, we can additionally get $u_a = 1$.

Proof. Let A, B, M and E be arbitrary positive constants and introduce new scaled variables:

$$\begin{aligned} \tilde{u} &= Au, & \tilde{t} &= Mt, \\ \tilde{\sigma} &= B\sigma, & \tilde{\boldsymbol{x}} &= E\boldsymbol{x}. \end{aligned}$$

Then

$$\begin{aligned} u_t &= \tfrac{M}{A}\tilde{u}_{\tilde{t}}, & \Delta u &= \tfrac{E^2}{A}\tilde{\Delta}\tilde{u}, \\ \sigma_t &= \tfrac{M}{B}\tilde{\sigma}_{\tilde{t}}, & \Delta \sigma &= \tfrac{E^2}{B}\tilde{\Delta}\tilde{\sigma}, \end{aligned}$$

where the Laplacian for the scaled space variable is given by

$$\tilde{\Delta} := \sum_{i=1}^{d} \left(\frac{\partial}{\partial \tilde{x}_i} \right)^2 .$$

Inserting these into (3.1a) and (3.1b), and then multiplying by $\frac{A}{M}$ (resp. $\frac{B}{M}$) gives

$$\begin{aligned} \tilde{u}_{\tilde{t}} - \frac{DE^2}{M}\tilde{\Delta}\tilde{u} - \frac{KE^2A}{MB}\tilde{\Delta}\tilde{\sigma} &= \frac{A}{M}f, \\ \tilde{\sigma}_{\tilde{t}} + \frac{\gamma(\tilde{u}/A)}{M}\tilde{\sigma} &= \frac{\mu\, B}{AM}\tilde{u} + \frac{\nu\, B}{A}\tilde{u}_{\tilde{t}} + \frac{B}{M}\hat{f}. \end{aligned}$$

As the coefficients do not contain a separate A but only the quotient B/A, we let $H = B/A$. Now, solving

$$\frac{DE^2}{M} = 1, \qquad \frac{KE^2}{MH} = 1, \qquad \frac{\mu\, H}{M} = 1,$$

for H, M and E, we obtain the following unique solution:

$$H = \frac{K}{D}, \qquad M = \frac{\mu\, K}{D}, \qquad E = \frac{\sqrt{\mu\, K}}{D}.$$

Note that, as the value of H is fixed by these, the coefficient containing ν cannot be scaled to ± 1 as well.

It remains to choose a value for A and B. We choose $A = 1/u_a$. Then $u = u_a$ implies $\tilde{u} = 1$, as required. For the value of B we get $B = AH = K/(Du_a)$.

Thus after omitting the tilde marks from the new equations and redefining μ, f, $\hat{f}$, g, γ, $\breve{u}$ and $\breve{\sigma}$, we get again the problem 3.1 but this time with $D = K = \mu = 1$. □

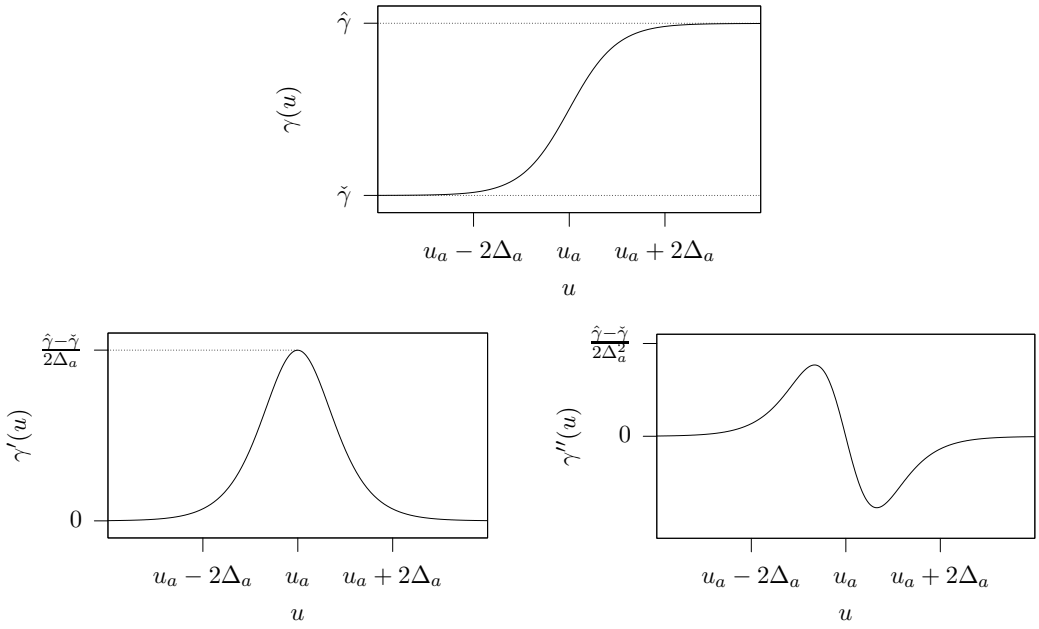

Figure 3.1: Graphs of $\gamma = \gamma_1$ and its derivatives. This is the tanh-version of γ.

3.5 The nonlinearity

3.5.1 Choices for nonlinearity

In this subsection we present three possible choices for the nonlinear function γ.

All these three functions have in common that they depend on constants $\check{\gamma}$ and $\hat{\gamma}$ giving the minimum and maximum value and on two further constants u_a and Δ_a.

The function value $\gamma(u)$ is close to $\check{\gamma}$ for small values of u and close to $\hat{\gamma}$ for large values of u. In between there is a transition region around $u = u_a$. The size of this transition region is proportional to Δ_a.

The first of the three choices for γ was introduced in [CW91] and was used in several other articles, e.g. [CWW95], [RS06].

Lemma 3.6 (Usual choice for γ).
Let $0 < \check{\gamma} \leq \hat{\gamma}$ and $u_a, \Delta_a > 0$. The specific $\gamma = \gamma_1$, which is given by

$$\gamma_1(u) = \frac{1}{2}(\hat{\gamma} + \check{\gamma}) + \frac{1}{2}(\hat{\gamma} - \check{\gamma}) \tanh\left(\frac{u - u_a}{\Delta_a}\right) \tag{3.5}$$

satisfies Assumption 3.4 with

$$C'_\gamma = \frac{\hat{\gamma} - \check{\gamma}}{2\Delta_a},$$
$$C''_\gamma = \frac{\hat{\gamma} - \check{\gamma}}{{\Delta_a}^2} \textit{(not optimal).}$$

Proof. See [RS05]. □

See Figure 3.1 for a graph of this standard version of γ. The above choice of γ has the property that the transition region is 'smeared out', the values of $\check{\gamma}$ and $\hat{\gamma}$ are attained only in the limit $u \to \pm\infty$. For certain theoretical results, (see [NLM$^+$]) it is more useful to have a transition region, which is limited, and with γ constant outside of this region. We now present two choices for function γ that have this property.

Lemma 3.7 (Another choice for γ).
Let $0 < \check{\gamma} \leq \hat{\gamma}$ and $u_a, \Delta_a > 0$ and define

$$p_\gamma(x) := \frac{3}{8}x^5 - \frac{5}{4}x^3 + \frac{15}{8}x.$$

Then if function $\gamma = \gamma_2$ is given by

$$\gamma_2(u) := \begin{cases} \check{\gamma} & \text{if } u \leq u_a - \frac{1}{2}\Delta_a, \\ \frac{\hat{\gamma}+\check{\gamma}}{2} + \frac{\hat{\gamma}-\check{\gamma}}{2} p_\gamma\left(\frac{u-u_a}{\frac{1}{2}\Delta_a}\right) & \text{if } |u-u_a| < \frac{1}{2}\Delta_a, \\ \hat{\gamma} & \text{if } u \geq u_a + \frac{1}{2}\Delta_a, \end{cases} \tag{3.6}$$

it satisfies Assumption 3.4 with

$$C'_\gamma = \frac{15(\hat{\gamma}-\check{\gamma})}{8\Delta_a},$$
$$C''_\gamma = \frac{10(\hat{\gamma}-\check{\gamma})}{\sqrt{3}{\Delta_a}^2}.$$

Proof. The derivatives of p_γ are

$$p'_\gamma(x) = \frac{15}{8}x^4 - \frac{15}{4}x^2 + \frac{15}{8} = \frac{15}{8}(x+1)^2(x-1)^2,$$
$$p''_\gamma(x) = \frac{15}{2}x^3 - \frac{15}{2}x = \frac{15}{2}(x+1)x(x-1).$$

Hence we have that,

$$p_\gamma(1) = 1,\ p_\gamma(-1) = -1,\ p'_\gamma(1) = p'_\gamma(-1) = p''_\gamma(1) = p''_\gamma(-1) = 0,$$

which implies that $\gamma \in C^2(\mathbb{R})$. Furthermore, we see that for $|u - u_a| < \frac{1}{2}\Delta_a$

$$\gamma'(u) = \frac{\hat{\gamma}-\check{\gamma}}{2(\frac{1}{2}\Delta_a)} p'_\gamma\left(\frac{u-u_a}{\frac{1}{2}\Delta_a}\right),$$

and

$$\gamma''(u) = \frac{\hat{\gamma}-\check{\gamma}}{2(\frac{1}{2}\Delta_a)^2} p''_\gamma\left(\frac{u-u_a}{\frac{1}{2}\Delta_a}\right).$$

Since $\|p'_\gamma\|_{L_\infty(-1,1)} = 15/8$ and $\|p''_\gamma\|_{L_\infty(-1,1)} \leq 15/2$, we get the given size of the constants. □

See Figure 3.2 for a graph of this second version of γ.

Lemma 3.8 (A third choice for γ).
Let $0 < \check{\gamma} \leq \hat{\gamma}$ and $u_a, \Delta_a > 0$ and define

$$f(x) := x + \frac{1}{\pi}\sin(\pi x).$$

Then if function $\gamma = \gamma_3$ is given by

$$\gamma_3(u) := \begin{cases} \check{\gamma} & \text{if } u \leq u_a - \frac{1}{2}\Delta_a, \\ \frac{\hat{\gamma}+\check{\gamma}}{2} + \frac{\hat{\gamma}-\check{\gamma}}{2} f\left(\frac{u-u_a}{\frac{1}{2}\Delta_a}\right) & \text{if } |u-u_a| < \frac{1}{2}\Delta_a, \\ \hat{\gamma} & \text{if } u \geq u_a + \frac{1}{2}\Delta_a, \end{cases} \tag{3.7}$$

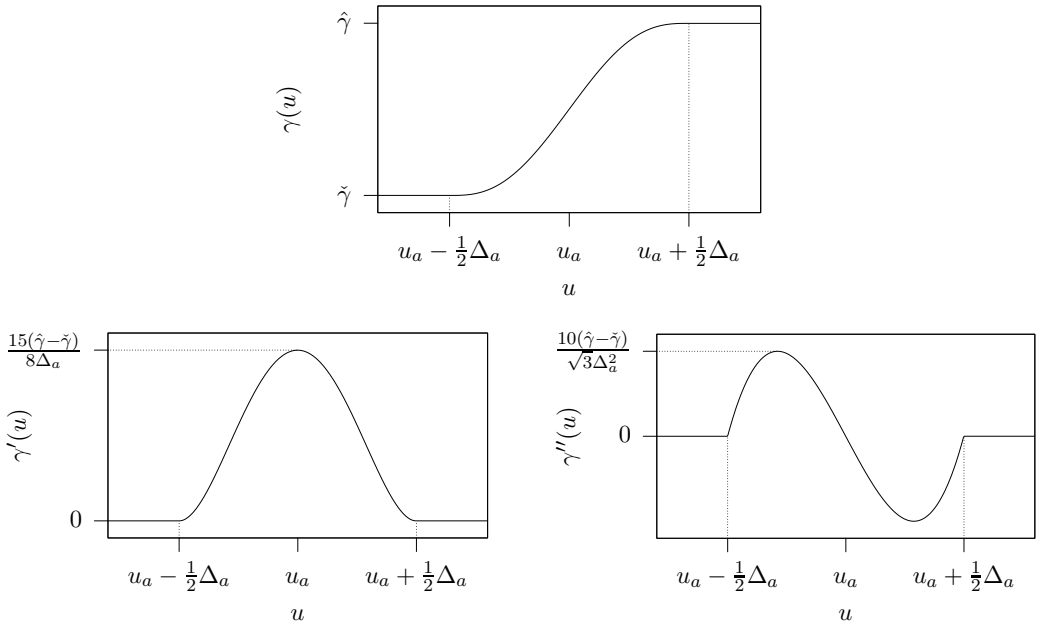

Figure 3.2: Graphs of $\gamma = \gamma_2$ and its derivatives. This is the polynomial version of γ.

it satisfies Assumption 3.4 with

$$C'_\gamma = \frac{2(\hat{\gamma} - \check{\gamma})}{\Delta_a},$$
$$C''_\gamma = \frac{2\pi \cdot (\hat{\gamma} - \check{\gamma})}{{\Delta_a}^2}.$$

Proof. The derivatives of f are

$$f'(x) = 1 + \cos(\pi x),$$
$$f''(x) = -\pi \sin(\pi x).$$

Hence we have that,

$$f(1) = 1,\ f(-1) = -1,\ f'(1) = f'(-1) = f''(1) = f''(-1) = 0,$$

which implies that $\gamma \in C^2(\mathbb{R})$. Furthermore, we see that for $|u - u_a| < \frac{1}{2}\Delta_a$

$$\gamma'(u) = \frac{\hat{\gamma} - \check{\gamma}}{2(\frac{1}{2}\Delta_a)} f'\left(\frac{u - u_a}{\frac{1}{2}\Delta_a}\right),$$

and

$$\gamma''(u) = \frac{\hat{\gamma} - \check{\gamma}}{2(\frac{1}{2}\Delta_a)^2} f''\left(\frac{u - u_a}{\frac{1}{2}\Delta_a}\right).$$

Since $\|f'\|_{L_\infty(-1,1)} = 2$ and $\|f''\|_{L_\infty(-1,1)} = \pi$, we get the given size of the constants. □

See Figure 3.3 for a graph of this third version of γ.

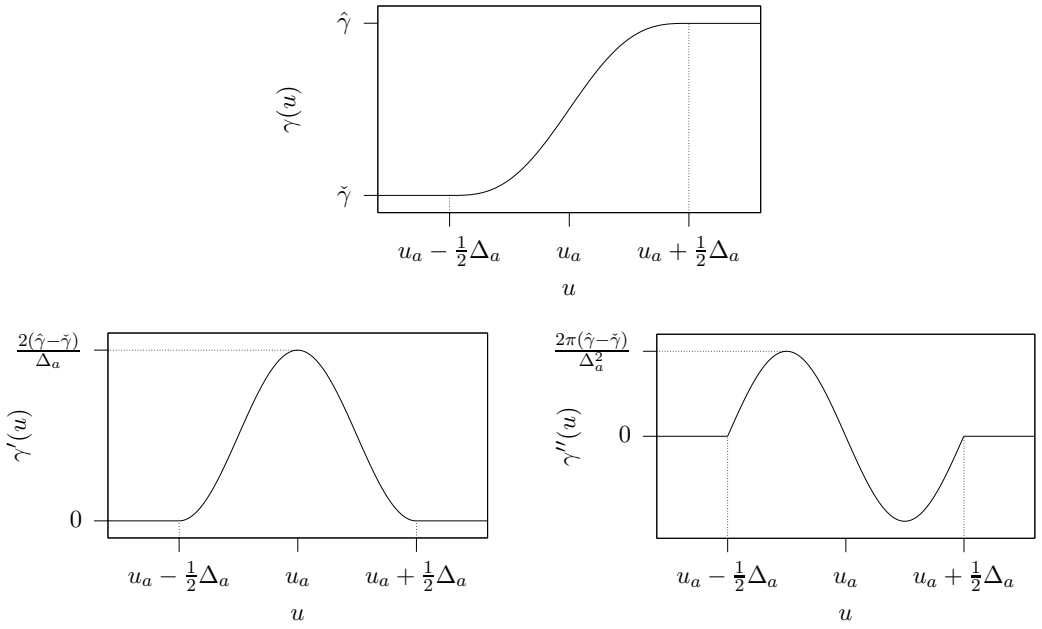

Figure 3.3: Graphs of $\gamma = \gamma_3$ and its derivatives. This is the $\sin(x) + x$-version of γ.

3.5.2 Basic properties of γ

In this subsection two small lemmas involving differences of function values of γ are given. These will be needed later for proving the error estimate. Lemma 3.9 provides a bound for

$$\|\gamma(v) - \gamma(w)\|_{L_p(\Omega)},$$

while Lemma 3.11 gives a bound for

$$\|\gamma(\tfrac{1}{2}v + \tfrac{1}{2}w) - \tfrac{1}{2}\gamma(v) - \tfrac{1}{2}\gamma(w)\|_{L_2(\Omega)}.$$

Lemma 3.9 (Difference of γ-values in L_p-norm).
Let $1 \le p \le \infty$ and $v, w \in L_p(\Omega)$. Then

$$\|\gamma(v) - \gamma(w)\|_{L_p(\Omega)} \le C'_\gamma \|v - w\|_{L_p(\Omega)}.$$

Proof. From Taylor's theorem we get that for all $a, b \in \mathbb{R}$ there is $\vartheta \in \mathbb{R}$ with either $a \le \vartheta \le b$ or $b \le \vartheta \le a$ such that

$$\gamma(b) - \gamma(a) = \gamma'(\vartheta)(b - a).$$

Thus, for any $a, b \in \mathbb{R}$,

$$|\gamma(b) - \gamma(a)| \le C'_\gamma |b - a|.$$

This can be used together with the definition of the L_p-norm to finish the proof. □

Before we state the second result of this subsection, Lemma 3.11, we prove the following lemma:

Lemma 3.10 (A bound for C^2-functions).
If $f : \mathbb{R} \to \mathbb{R}$ is twice continuously differentiable with bounded second derivative, $|f''(x)| \le \alpha \; \forall \, x \in \mathbb{R}$, then

$$|f(\tfrac{1}{2}a + \tfrac{1}{2}b) - \tfrac{1}{2}f(a) - \tfrac{1}{2}f(b)| \le \frac{\alpha}{8} |b - a|^2,$$

for all $a, b \in \mathbb{R}$.

Proof. From Taylor's theorem we get

$$f(a) = f(\tfrac{a+b}{2}) + \tfrac{a-b}{2} f'(\tfrac{a+b}{2}) + \tfrac{1}{2}\big(\tfrac{a-b}{2}\big)^2 f''(\xi_1),$$
$$f(b) = f(\tfrac{a+b}{2}) + \tfrac{b-a}{2} f'(\tfrac{a+b}{2}) + \tfrac{1}{2}\big(\tfrac{b-a}{2}\big)^2 f''(\xi_2),$$

for some ξ_1 between a and $\frac{a+b}{2}$, and some ξ_2 between b and $\frac{a+b}{2}$. Adding these gives

$$f(a) + f(b) = 2f(\tfrac{a+b}{2}) + \tfrac{1}{2}\big(\tfrac{a-b}{2}\big)^2 \big(f''(\xi_1) + f''(\xi_2)\big),$$

which can be rearranged to

$$f(\tfrac{a+b}{2}) - \tfrac{1}{2} f(a) - \tfrac{1}{2} f(b) = \frac{(b-a)^2}{16}(f''(\xi_1) + f''(\xi_2)).$$

We conclude the proof by observing that $|f''(\xi_1) + f''(\xi_2)| \le 2\alpha$. □

Lemma 3.11 (Another bound for γ-values).
Let $v, w \in L_4(\Omega)$. Then

$$\|\gamma(\tfrac{1}{2}v + \tfrac{1}{2}w) - \tfrac{1}{2}\gamma(v) - \tfrac{1}{2}\gamma(w)\|_{L_2(\Omega)} \le \frac{C''_\gamma}{8}\|v - w\|^2_{L_4(\Omega)}.$$

Proof. Apply Lemma 3.10 to function γ, using $|\gamma''(x)| \le C''_\gamma$ from Assumption 3.4.

$$\begin{aligned}&\|\gamma(\tfrac{1}{2}v + \tfrac{1}{2}w) - \tfrac{1}{2}\gamma(v) - \tfrac{1}{2}\gamma(w)\|_{L_2(\Omega)} \\ &\quad \le \|\tfrac{C''_\gamma}{8}(v-w)^2\|_{L_2(\Omega)} = \tfrac{C''_\gamma}{8}\|(v-w)^2\|_{L_2(\Omega)} = \tfrac{C''_\gamma}{8}\|v-w\|^2_{L_4(\Omega)},\end{aligned}$$

where we have used that $\|a^2\|_{L_2(\Omega)} = \|a\|^2_{L_4(\Omega)}$ for any $a \in L_4(\Omega)$. □

Chapter 4

Existence and Uniqueness

This chapter deals with questions of existence and uniqueness of solutions and their stability.

As is usual for the finite element analysis, the first step is to give a variational formulation of the problem, since the finite element functions we use do not have a classical second space derivative. This variational formulation will be given in Section 4.1.

Then, in Section 4.2 we present a bound for the exact solution of the variational formulation in terms of the problem data. This is a first step towards establishing stability of the problem.

Section 4.3 gives a result of existence and uniqueness of a linearised version of the problem, where the term $\gamma(u)$ in (3.1b) is replaced by a time- and space-dependent function $\tilde{\gamma}$, which is independent of u.

In Section 4.4 the main result of this chapter is established: In the case of 1-dimensional domains, the problem has a unique solution in our variational formulation.

4.1 Variational formulation

The following equations can be derived from Problem 3.1 with $\nu = 0$. Taking the L_2 inner product of (3.1a) with a function $v \in V$ and using Green's formula (Lemma 2.8) gives (4.1a), provided the functions are sufficiently regular.

Problem 4.1 (First variational formulation of the Cohen model).
Suppose we are given $u_D \in C^1([0,T]; H^1(\Omega))$, $f \in L_2(0,T;V')$, $g \in L_2(0,T;L_2(\Gamma_N))$ and $\hat{f} \in C([0,T]; L_2(\Omega))$, then find $u \in u_D + \mathcal{W}(0,T;V,V')$ and $\sigma \in L_2(0,T;H^1(\Omega)) \cap C^1([0,T];L_2(\Omega))$ such that

$$\langle \dot{u}, v\rangle + (D\nabla u, \nabla v) + (K\nabla\sigma, \nabla v) = \langle L, v\rangle \qquad \forall\, v \in V, \text{ a.e. } t \in (0,T), \tag{4.1a}$$

$$\dot{\sigma} + \gamma(u)\sigma = \mu\, u + \hat{f} \qquad \forall\, t \in [0,T], \tag{4.1b}$$

$$u(0) = \breve{u}, \tag{4.1c}$$

$$\sigma(0) = \breve{\sigma}, \tag{4.1d}$$

where $L \in L_2(0,T;V')$ is defined by

$$\langle L, v\rangle := (f,v)_0 + (g,v)_{\Gamma_N}. \tag{4.2}$$

For most of the following boundedness results and error analysis the variational formulation given by Problem 4.1 is inconvenient for two reasons:

- Firstly, carrying around the problem parameters D, K, μ would be awkward. Therefore, we rescale the problem using Lemma 3.5 and we can set $D = K = \mu = 1$.
- Secondly, for the results of this chapter and later for the error estimate, we would like that both u and σ are contained in the space $L_2(0,T;V)$. Since the space V enforces homogeneous Dirichlet boundary conditions, we are lead to the following assumption:

Assumption 4.2 (Homogeneous Dirichlet boundary conditions).
We assume that,

- the Dirichlet boundary condition on u is homogeneous: $u_D \equiv 0$,
- σ vanishes on the Dirichlet boundary: $\sigma|_{\Gamma_D} = 0$.

The second of these assumptions, $\sigma|_{\Gamma_D} = 0$, can also be derived under certain conditions: If we assume that $u_D = \breve{\sigma}|_{\Gamma_D} = \hat{f}|_{\Gamma_D} = 0$ and provided that u and σ are sufficiently regular, then (3.1b) implies that $\sigma|_{\Gamma_D} = 0$. Thus, the assumption that σ vanishes on the Dirichlet boundary is not as far-fetched as it might appear.

For the existence and uniqueness results, we only need that $u(t) + \sigma(t) \in V$. The requirement that both $u(t)$ and $\sigma(t)$ are contained in V will be needed for the error estimate in Chapter 6.

We will consider the following variational formulation as the basis for the boundedness and error analysis instead of Problem 4.1:

Problem 4.3 (Simplified variational formulation of the Cohen model).
Given $\breve{u}, \breve{\sigma} \in L_2(\Omega)$ and $L \in L_2(0,T;V')$ and $\hat{f} \in C([0,T];L_2(\Omega))$, find $(u,\sigma) \in C([0,T];L_2(\Omega)^2)$ with $\sigma \in C^1([0,T];L_2(\Omega))$, $u + \sigma \in L_2(0,T;V)$ and $\dot{u} \in L_2(0,T;V')$, such that

$$\langle \dot{u}, v\rangle + (u+\sigma, v)_V = \langle L, v\rangle \quad \forall\, v \in V, \text{ a.e. } t \in (0,T), \tag{4.3a}$$

$$\dot{\sigma} + \gamma(u)\sigma = u + \hat{f} \quad t \in [0,T], \tag{4.3b}$$

$$u(0) = \breve{u}, \tag{4.3c}$$

$$\sigma(0) = \breve{\sigma}. \tag{4.3d}$$

Taking the L_2 inner product of both sides of (4.3b) with an arbitrary $v \in L_2(\Omega)$ gives:

$$(\dot{\sigma}, v)_0 + (\gamma(u)\sigma, v)_0 = (u, v)_0 + (\hat{f}, v)_0 \qquad \forall\, v \in L_2(\Omega), t \in [0,T]. \tag{4.4}$$

Mostly we will use this equality instead of (4.3b). Note however that in Lemma 4.16 we will need an estimate that is based on the "pointwise" equality (4.3b) instead of the variational form (4.4). This is the reason why (4.3b) is formulated in a non-variational form.

4.1.1 Two approaches for getting results

One of the problems in dealing with the equations, either to prove stability, discrete stability or an error estimate, lies in the term $(\sigma, v)_V$ of the first equation (and its discrete counterpart). It seems that there are two ways of handling this term:

- Using the inverse Laplacian to cancel it with the term $(u, v)_0$ from the second equation
$$(\sigma, \mathcal{G}u)_V - (u, \sigma)_0 = 0.$$
The details of this can be seen in the proof of Proposition 4.5.

- Adding the second equation (4.4) to the first one (4.3a) and working with the sum $u + \sigma$ instead of only u:

$$(\dot{u} + \dot{\sigma}, v)_0 + (u + \sigma, v)_V + (\gamma(u)\sigma, v)_0 = (u, v)_0 + (\hat{f}, v)_0 + \langle L, v\rangle \quad \forall v \in V. \tag{4.5}$$

  It appears that $u + \sigma$ has a higher regularity than either u or σ on their own. For the existence and uniqueness results of this chapter, we will use (4.5) together with (4.4).

The next lemma shows that it is equivalent to have bounds for either

- $\|u\|$ and $\|\sigma\|$, or
- $\|u + \sigma\|$ and $\|\sigma\|$.

This will be needed for the second approach.

Lemma 4.4 (Consequences of triangle inequality).
For any normed vector space $(X, \|\cdot\|)$ and any $a, b \in X$, we have that

$$\|a + b\|^2 \leq 2(\|a\|^2 + \|b\|^2), \tag{4.6}$$

and

$$\frac{1}{3}(\|a\|^2 + \|b\|^2) \leq \|a + b\|^2 + \|b\|^2 \leq 3(\|a\|^2 + \|b\|^2). \tag{4.7}$$

Proof. Firstly, $\|a+b\|^2 \leq (\|a\| + \|b\|)^2 \leq \|a\|^2 + 2\|a\|\|b\| + \|b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$. Thus $\|a + b\|^2 + \|b\|^2 \leq 2\|a\|^2 + 3\|b\|^2$, which proves the right part of the second inequality. To get the left part of the second inequality put $a = a' + b'$ and $b = -b'$ into the right part. □

4.2 Boundedness of exact solutions

We give two similar results of basic stability, i.e. of bounds on the exact solution in different norms. The reason for giving both results is to show the use of both approaches mentioned in the previous section.

4.2.1 Using the inverse Laplacian

For the first bound on the exact solution we use the inverse Laplacian.

Proposition 4.5 (Basic stability in weak norms).
If (u, σ) is a solution to Problem 4.3, then

$$\left.\begin{array}{c} \|u\|_{C([0,T];V')} \\ \|\sigma\|_{C([0,T];L_2(\Omega))} \\ \|u\|_{L_2(0,T;L_2(\Omega))} \\ \check{\gamma}^{1/2}\|\sigma\|_{L_2(0,T;L_2(\Omega))} \end{array}\right\} \leq \begin{array}{r} \|\breve{u}\|_{V'} + \|\breve{\sigma}\|_0 + C_V\|L\|_{L_2(0,T;V')} \\ + \check{\gamma}^{-1/2}\|\hat{f}\|_{L_2(0,T;L_2(\Omega))}. \end{array} \tag{4.8}$$

Proof. Choose $v = \mathcal{G}u$ in (4.3a) and use the properties of the inverse Laplacian (2.16) and (2.25),

$$\frac{1}{2}\frac{d}{dt}\|u\|_{V'}^2 + \|u\|_0^2 + (\sigma, u)_0 = \langle L, \mathcal{G}u\rangle.$$

Choose $v = \sigma$ in (4.4) and use (2.21),

$$\frac{1}{2}\frac{d}{dt}\|\sigma\|_0^2 + (\gamma(u)\sigma, \sigma)_0 = (u, \sigma)_0 + (\hat{f}, \sigma)_0.$$

Now adding these two equations eliminates the cross terms $(\sigma, u)_0$ and we get, after multiplying by 2,

$$\frac{d}{dt}\|u\|_{V'}^2 + \frac{d}{dt}\|\sigma\|_0^2 + 2\|u\|_0^2 + 2(\gamma(u)\sigma, \sigma)_0 = 2\langle L, \mathcal{G}u\rangle + 2(\hat{f}, \sigma)_0. \tag{4.9}$$

From the definition of the L_2 inner product, we see that $(\gamma(u)\sigma, \sigma)_0 \geq \check{\gamma}\|\sigma\|_0^2$. Also, using the definition of the dual norm $\|\cdot\|_{V'}$, the inverse Laplacian property (2.17), then (2.14) to weaken the norm and finally Young's inequality (A.1), we get

$$\langle L, \mathcal{G}u\rangle \leq \|L\|_{V'} \cdot \|\mathcal{G}u\|_V = \|L\|_{V'} \cdot \|u\|_{V'} \leq \|L\|_{V'} \cdot C_V\|u\|_0 \leq \frac{C_V^2}{2}\|L\|_{V'}^2 + \frac{1}{2}\|u\|_0^2.$$

Furthermore, the Cauchy-Schwarz inequality and Young's inequality yield

$$(\hat{f}, \sigma)_0 \leq \|\hat{f}\|_0 \cdot \|\sigma\|_0 \leq \frac{1}{2\check{\gamma}}\|\hat{f}\|_0^2 + \frac{\check{\gamma}}{2}\|\sigma\|_0^2.$$

Using these three inequalities in (4.9) and subtracting $\|u\|_0^2$ and $\check{\gamma}\|\sigma\|_0^2$ from both sides yields

$$\frac{d}{dt}\|u\|_{V'}^2 + \frac{d}{dt}\|\sigma\|_0^2 + \|u\|_0^2 + \check{\gamma}\|\sigma\|_0^2 \leq C_V^2\|L\|_{V'}^2 + \tfrac{1}{\check{\gamma}}\|\hat{f}\|_0^2.$$

Integrating over $(0, t)$ and using the initial conditions (4.3c) and (4.3d) gives

$$\|u(t)\|_{V'}^2 - \|\breve{u}\|_{V'}^2 + \|\sigma(t)\|_0^2 - \|\breve{\sigma}\|_0^2 + \int_0^t \|u(s)\|_0^2\, ds + \check{\gamma}\int_0^t \|\sigma(s)\|_0^2\, ds \leq C_V^2\int_0^t \|L(s)\|_{V'}^2\, ds + \frac{1}{\check{\gamma}}\int_0^t \|\hat{f}(s)\|_0^2\, ds.$$

We rearrange and use the definition of the norms $L_2(0, t; \cdot)$:

$$\|u(t)\|_{V'}^2 + \|\sigma(t)\|_0^2 + \|u\|_{L_2(0,t;L_2(\Omega))}^2 + \check{\gamma}\|\sigma\|_{L_2(0,t;L_2(\Omega))}^2 \leq \|\breve{u}\|_{V'}^2 + \|\breve{\sigma}\|_0^2 + C_V^2\|L\|_{L_2(0,t;V')}^2 + \tfrac{1}{\check{\gamma}}\|\hat{f}\|_{L_2(0,t;L_2(\Omega))}^2. \tag{4.10}$$

Now observe that by (A.8) for $p = 2$ and $n = 4$, we have

$$(\|\breve{u}\|_{V'}^2 + \|\breve{\sigma}\|_0^2 + C_V^2\|L\|_{L_2(0,t;V')}^2 + \tfrac{1}{\check{\gamma}}\|\hat{f}\|_{L_2(0,t;L_2(\Omega))}^2)^{1/2} \leq \|\breve{u}\|_{V'} + \|\breve{\sigma}\|_0 + C_V\|L\|_{L_2(0,t;V')} + \check{\gamma}^{-1/2}\|\hat{f}\|_{L_2(0,t;L_2(\Omega))}. \tag{4.11}$$

Thus we can look at each of the four terms on the left-hand side of (4.10) and take the square root to obtain (4.8). □

4.2.2 Using the $u + \sigma$-approach

With the $u + \sigma$-approach, we can obtain another bound on the exact solution, which is similar to Proposition 4.5. Here, however the norms are stronger. Note that $u + \sigma$ appears in the V-norm, indicating that $u + \sigma$ might be smoother in space than u and σ are. Later in Chapter 7 we will see numerical simulations that support this conjecture.

Proposition 4.6 (Basic stability in stronger norms).
If (u, σ) is a solution to Problem 4.3, then

$$\left.\begin{aligned}&\|u\|_{C([0,T];L_2(\Omega))}\\&\|\sigma\|_{C([0,T];L_2(\Omega))}\\&\|u\|_{L_2(0,T;L_2(\Omega))}\\&\|\sigma\|_{L_2(0,T;L_2(\Omega))}\\&\|u+\sigma\|_{L_2(0,T;V)}\end{aligned}\right\} \leq C\Big(\|\breve{u}\|_0 + \|\breve{\sigma}\|_0 + \|L\|_{L_2(0,T;V')} + \|\hat{f}\|_{L_2(0,T;L_2(\Omega))}\Big). \tag{4.12}$$

Proof. Add (4.3a) and (4.4) and choose $v = u + \sigma$, which yields, using (2.21),

$$\frac{1}{2}\frac{d}{dt}\|u+\sigma\|_0^2 + \|u+\sigma\|_V^2 + (\gamma(u)\sigma, u+\sigma)_0 = (u, u+\sigma)_0 + (\hat{f}, u+\sigma)_0 + \langle L, u+\sigma\rangle.$$

Bringing the term $(\gamma(u)\sigma, u)_0$ to the other side then integrating over $(0,t)$ and multiplying by 2 gives

$$\begin{aligned}&\|u(t)+\sigma(t)\|_0^2 + 2\int_0^t \|u(s)+\sigma(s)\|_V^2\,ds + 2\int_0^t (\gamma(u(s))\sigma(s), \sigma(s))_0\,ds\\&\quad= \|\breve{u}+\breve{\sigma}\|_0^2 - 2\int_0^t (\gamma(u(s))\sigma(s), u(s))_0\,ds + 2\int_0^t (u(s), u(s)+\sigma(s))_0\,ds\\&\qquad+ 2\int_0^t (\hat{f}(s), u(s)+\sigma(s))_0\,ds + 2\int_0^t \langle L(s), u(s)+\sigma(s)\rangle\,ds,\end{aligned}$$

where we have used the initial conditions (4.3c) and (4.3d). Now we use the fact that $(\gamma(u)\sigma, \sigma)_0 \geq \check{\gamma}\|\sigma\|_0^2$ and also that

$$(\gamma(u)\sigma, u)_0 \leq \|\gamma(u)\sigma\|_0 \cdot \|u\|_0 \leq \hat{\gamma}\|\sigma\|_0 \cdot \|u\|_0$$

and obtain

$$\begin{aligned}&\|u+\sigma\|_0^2 + 2\int_0^t \|u+\sigma\|_V^2\,ds + 2\check{\gamma}\int_0^t \|\sigma\|_0^2\,ds\\&\quad\leq \|\breve{u}+\breve{\sigma}\|_0^2 + 2\hat{\gamma}\int_0^t \|\sigma\|_0\|u\|_0\,ds + 2\int_0^t \|u\|_{V'}\|u+\sigma\|_V\,ds\\&\qquad+ 2\int_0^t \|\hat{f}\|_{V'}\|u+\sigma\|_V\,ds + 2\int_0^t \|L\|_{V'}\|u+\sigma\|_V\,ds.\end{aligned}$$

After several applications of Young's inequality (A.1) we get

$$\begin{aligned}&\|u+\sigma\|_0^2 + \int_0^t \|u+\sigma\|_V^2\,ds + \check{\gamma}\int_0^t \|\sigma\|_0^2\,ds\\&\quad\leq \|\breve{u}+\breve{\sigma}\|_0^2 + \frac{\hat{\gamma}^2}{\check{\gamma}}\int_0^t \|u\|_0^2\,ds + 3\int_0^t \|u\|_{V'}^2\,ds + 3\int_0^t \|\hat{f}\|_{V'}^2\,ds + 3\int_0^t \|L\|_{V'}^2\,ds.\end{aligned}$$

To change the norm from $\|\cdot\|_{V'}$ to $\|\cdot\|_0$ on the right-hand side, we use (2.14):

$$\begin{aligned}&\|u+\sigma\|_0^2 + \int_0^t \|u+\sigma\|_V^2\,ds + \check{\gamma}\int_0^t \|\sigma\|_0^2\,ds\\&\quad\leq \|\breve{u}+\breve{\sigma}\|_0^2 + \left(\frac{\hat{\gamma}^2}{\check{\gamma}} + 3C_V^2\right)\int_0^t \|u\|_0^2\,ds + 3C_V^2\int_0^t \|\hat{f}\|_0^2\,ds + 3\int_0^t \|L\|_{V'}^2\,ds.\end{aligned} \tag{4.13}$$

Now we would like to get a second inequality with which we can eliminate the term $\int_0^t \|u\|_0^2 ds$ on the right-hand side. We start again with (4.4) and choose $v = \sigma$ which gives us

$$\frac{1}{2}\frac{d}{dt}\|\sigma\|_0^2 + (\gamma(u)\sigma, \sigma)_0 = (u, \sigma)_0 + (\hat{f}, \sigma)_0.$$

Integrating and multiplying by 2 gives

$$\|\sigma(t)\|_0^2 + 2\int_0^t (\gamma(u(s))\sigma(s), \sigma(s))_0 \, ds \\ = \|\breve{\sigma}\|_0^2 + 2\int_0^t (u(s), \sigma(s))_0 \, ds + 2\int_0^t (\hat{f}(s), \sigma(s))_0 \, ds.$$

Using the Cauchy-Schwarz inequality:

$$\|\sigma(t)\|_0^2 + 2\check{\gamma}\int_0^t \|\sigma(s)\|_0^2 \, ds \leq \|\breve{\sigma}\|_0^2 + 2\int_0^t \|u(s)\|_0 \|\sigma(s)\|_0 \, ds + 2\int_0^t \|\hat{f}(s)\|_0 \|\sigma(s)\|_0 \, ds.$$

Using Young's inequality:

$$\|\sigma(t)\|_0^2 + \check{\gamma}\int_0^t \|\sigma(s)\|_0^2 \, ds \leq \|\breve{\sigma}\|_0^2 + \frac{2}{\check{\gamma}}\int_0^t \|u(s)\|_0^2 \, ds + \frac{2}{\check{\gamma}}\int_0^t \|\hat{f}(s)\|_0^2 \, ds. \tag{4.14}$$

Now, adding (4.13) and (4.14) yields:

$$\|u(t) + \sigma(t)\|_0^2 + \|\sigma(t)\|_0^2 + \int_0^t \|u(s) + \sigma(s)\|_V^2 \, ds + 2\check{\gamma}\int_0^t \|\sigma(s)\|_0^2 \, ds \\ \leq \|\breve{u} + \breve{\sigma}\|_0^2 + \|\breve{\sigma}\|_0^2 + 3\int_0^t \|L(s)\|_{V'}^2 \, ds + (3C_V^2 + \tfrac{2}{\check{\gamma}})\int_0^t \|\hat{f}(s)\|_0^2 \, ds \\ + \left(\frac{2 + \hat{\gamma}^2}{\check{\gamma}} + 3C_V^2\right)\int_0^t \|u(s)\|_0^2 \, ds.$$

Using (4.7) to "split" $\|u + \sigma\|_0$ and adding $\int_0^t \|u(s)\|_0^2 \, ds$ to both sides yields

$$\tfrac{1}{3}\|u(t)\|_0^2 + \tfrac{1}{3}\|\sigma(t)\|_0^2 + \int_0^t \|u(s) + \sigma(s)\|_V^2 \, ds + \int_0^t \|u(s)\|_0^2 \, ds + 2\check{\gamma}\int_0^t \|\sigma(s)\|_0^2 \, ds \\ \leq 3\|\breve{u}\|_0^2 + 3\|\breve{\sigma}\|_0^2 + 3\int_0^t \|L(s)\|_{V'}^2 \, ds + (3C_V^2 + \tfrac{2}{\check{\gamma}})\int_0^t \|\hat{f}(s)\|_0^2 \, ds \\ + \left(\frac{2 + \hat{\gamma}^2}{\check{\gamma}} + 3C_V^2 + 1\right)\int_0^t \|u(s)\|_0^2 \, ds.$$

We can now use Grönwall's lemma (Lemma (A.3)) to eliminate the $\|u\|_0^2$-term on the right-hand side. Also using the norm-notation instead of the integrals, we obtain

$$\begin{aligned} &\tfrac{1}{3}\|u(t)\|_0^2 + \tfrac{1}{3}\|\sigma(t)\|_0^2 + \|u + \sigma\|_{L_2(0,t;V)}^2 + \|u\|_{L_2(0,t;L_2(\Omega))}^2 + 2\check{\gamma}\|\sigma\|_{L_2(0,t;L_2(\Omega))}^2 \\ &\quad \leq c_1\left(\|\breve{u}\|_0^2 + \|\breve{\sigma}\|_0^2 + \|L\|_{L_2(0,t;V')}^2\right) + c_2\|\hat{f}\|_{L_2(0,t;L_2(\Omega))}^2, \end{aligned} \tag{4.15}$$

where $c_1 = 3\exp(3T((2+\hat{\gamma}^2)/\check{\gamma} + 3C_V^2 + 1))$ and $c_2 = c_1(C_V^2 + 2/(3\check{\gamma}))$. As a final step, looking separately at each of the terms on the right-hand side and taking the square root gives (4.12). □

4.3 Unique solution of linearised problem

This section shows that the linearised version of the problem has a unique solution. The proof uses Banach's fixed point theorem, similar to the way it is used in [Eva98, Subsection 9.2.1].

In order to use this fixed point theorem we have to show that a certain mapping is a contraction. As this does not work using the standard norm of the space involved, we have to introduce so called *weighted norms* to get the contraction property. In Subsection 4.3.1 these norms are defined and a lemma is proved, which will be used to show the contraction property. Then, in Subsection 4.3.2 we will prove the uniqueness of the solution to the linearised problem by using weighted norms.

4.3.1 Weighted norms

In this subsection we describe a trick that is used in some standard analysis textbooks to prove the Picard-Lindelöf theorem (a theorem that assures unique solutions to certain initial value problems). The problem is reformulated as an integral equation and an iteration is set up, the *Picard iteration*. To show that the sequence of iterates converges, a so called *weighted norm* is used. This makes it possible to use Banach's fixed point theorem A.5. This approach is used for example in [Hac95, 2.1.1] and in [Heu00, exercise A117.3].

We emphasise that the idea of using weighted norms as described in this subsection is not new, but merely an expanded and modified version of the idea which appears in [Hac95] and [Heu00].

Definition of the weighted norms

The space $C([a,b];X)$ can also be equipped with norms other than the standard norm from (2.20). These other norms will turn out to be equivalent.

Definition 4.7 (Weighted max**-norm).**
For $L \in \mathbb{R}$ and $y \in C([a,b];X)$ define a *weighted* max*-norm* by

$$\|y\|_{C([a,b];X),L} := \max_{t\in[a,b]} e^{-L(t-a)}\|y(t)\|_X. \tag{4.16}$$

For a more general definition, see [GLS90, Section 4.2].

Lemma 4.8 (Equivalence of weighted max**-norms).**
For $L \in \mathbb{R}$, $\|\cdot\|_{C([a,b];X),L}$ is a norm on the space $C([a,b];X)$. Moreover, all the norms $\|\cdot\|_{C([a,b];X),L}$ are equivalent to each other and to $\|\cdot\|_{C([a,b];X)}$.

Proof. The norm axioms can be verified immediately from the definition. Also observe that $\|\cdot\|_{C([a,b];X),0} = \|\cdot\|_{C([a,b];X)}$. Now, let $L, M \in \mathbb{R}$. Then for any $y \in C([a,b];X)$

$$\begin{aligned}
\|y\|_{C([a,b];X),L} &= \max_{t\in[a,b]} e^{-L(t-a)}\|y(t)\|_X \\
&= \max_{t\in[a,b]} e^{(M-L)(t-a)}e^{-M(t-a)}\|y(t)\|_X \\
&\le \max(1, e^{(M-L)(b-a)}) \cdot \max_{t\in[a,b]} e^{-M(t-a)}\|y(t)\|_X \\
&= \max(1, e^{(M-L)(b-a)}) \cdot \|y\|_{C([a,b];X),M}.
\end{aligned}$$

This bounds the norm $\|y\|_{C([a,b];X),L}$ in terms of the norm $\|y\|_{C([a,b];X),M}$. Exchanging the roles of L and M yields the bound in the other direction. □

Showing the contraction property

We now come to the lemma, that will be used in two proofs of existence and uniqueness of solutions:

- in Proposition 4.10 for a linearised problem, and
- in Proposition 4.18 for the nonlinear problem in 1D.

Lemma 4.9 (Trick with weighted max**-norm).**
Let $v, w \in C([a,b];X)$ and $p \geq 1$. Suppose that there are constants $L > 0$ and $M \geq 0$ such that the inequality

$$\|v(t)\|_X^p \leq M^p + L \int_a^t \|w(s)\|_X^p \, ds$$

holds for $t \in [a,b]$. Then, for any $q > 0$ there is a weighted norm $\|\cdot\|_{C([a,b];X),\beta}$ such that

$$\|v\|_{C([a,b];X),\beta} \leq M + q \, \|w\|_{C([a,b];X),\beta}.$$

In fact, we can choose $\beta = Lq^{-p}/p$.

Proof. To begin with, we only assume $\beta > 0$, the correct value will be chosen later. From the assumption, we have

$$\begin{aligned}
\|v(t)\|_X^p &\leq M^p + L \int_a^t e^{p\beta(s-a)} \, \left(e^{-\beta(s-a)} \|w(s)\|_X\right)^p \, ds \\
&\leq M^p + L \int_a^t e^{p\beta(s-a)} \|w\|_{C([a,t];X),\beta}^p \, ds \\
&= M^p + L \|w\|_{C([a,t];X),\beta}^p \int_a^t e^{p\beta(s-a)} \, ds \\
&= M^p + L \, \|w\|_{C([a,t];X),\beta}^p \, \frac{1}{p\,\beta} (e^{p\beta(t-a)} - 1) \\
&\leq M^p + \frac{L}{p\,\beta} \, \|w\|_{C([a,t];X),\beta}^p \, e^{p\beta(t-a)}.
\end{aligned}$$

Now multiplying by $e^{-p\beta(t-a)}$ gives

$$e^{-p\beta(t-a)} \|v(t)\|_X^p \leq e^{-p\beta(t-a)} M^p + \frac{L}{p\,\beta} \|w\|_{C([a,b];X),\beta}^p,$$

taking the p-th root and using Corollary A.12,

$$e^{-\beta(t-a)} \|v(t)\|_X \leq e^{-\beta(t-a)} M + \left(\frac{L}{p\,\beta}\right)^{1/p} \|w\|_{C([a,b];X),\beta}.$$

This implies, since $\max_{t\in[a,b]} e^{-\beta(t-a)} = 1$,

$$e^{-\beta(t-a)} \|v(t)\|_X \leq M + \left(\frac{L}{p\,\beta}\right)^{1/p} \|w\|_{C([a,b];X),\beta}.$$

The right-hand side is now independent of t, thus, by taking the maximum on the left-hand side, we get

$$\|v\|_{C([a,b];X),\beta} \leq M + \left(\frac{L}{p\,\beta}\right)^{1/p} \|w\|_{C([a,b];X),\beta}.$$

Choosing $\beta = Lq^{-p}/p$ concludes the proof. □

4.3.2 The result

In this subsection we prove the existence and uniqueness of the linearised problem. In the proof we set up an iteration, use weighted norms to show that it is a contraction and Banach's fixed point theorem to obtain a fixed point.

Proposition 4.10 (Unique solution of linearised Cohen model).
If we replace $\gamma(u)$ in Problem 4.3 by a time- and space-dependent function

$$\tilde{\gamma} \in C([0,T];L_\infty(\Omega)),$$

that also satisfies $0 < \check{\gamma} \leq \tilde{\gamma}(t,\boldsymbol{x}) \leq \hat{\gamma}$, then the problem has a unique solution

$$(u,\sigma) \in C([0,T];L_2(\Omega)^2)$$

with $u+\sigma \in \mathcal{W}(0,T;V,V')$ and $\sigma \in C^1([0,T];L_2(\Omega))$.

Remark 4.11.
Here it is necessary to point out how to interpret the equations, since $u \in C([0,T];L_2(\Omega))$ does not imply that $\dot{u}$ exists. But $u+\sigma \in \mathcal{W}(0,T;V,V')$ implies that $\frac{\partial}{\partial t}(u+\sigma) \in L_2(0,T;V')$ and since $\dot{\sigma} \in C([0,T];L_2(\Omega))$, we can write

$$\dot{u} = \frac{\partial}{\partial t}(u+\sigma) - \dot{\sigma}$$

and thus obtain $\dot{u} \in L_2(0,T;V')$.

Proof of Proposition 4.10. We set $y := u+\sigma$ and reformulate the problem in terms of (y,σ) instead of (u,σ). Add (4.4) to (4.3a) to get a first equation and use (4.3b) as a second equation, then replace u by $y-\sigma$ in both to obtain the following problem:
Find $y \in \mathcal{W}(0,T;V,V')$ and $\sigma \in C^1([0,T];L_2(\Omega))$ such that

$$\langle \dot{y}, v\rangle + (y,v)_V = (y-(1+\tilde{\gamma})\sigma, v)_0 + (\hat{f},v)_0 + \langle L, v\rangle, \tag{4.17a}$$
$$\dot{\sigma} + (1+\tilde{\gamma})\sigma = y + \hat{f}, \tag{4.17b}$$
$$y(0) = \breve{u} + \breve{\sigma}, \tag{4.17c}$$
$$\sigma(0) = \breve{\sigma}. \tag{4.17d}$$

Now, (4.17) is equivalent to the linearised form of Problem 4.3. We proceed by showing that (4.17) has a unique solution. To this end, we set up a mapping

$$A : C([0,T];L_2(\Omega)^2) \to C([0,T];L_2(\Omega)^2).$$

We set $(y,\sigma) = A(y^*,\sigma^*)$, where the image (y,σ) is defined as the solution of the following system, using the pre-image $(y^*,\sigma^*) \in C([0,T];L_2(\Omega)^2)$,

$$(\dot{y}, v)_0 + (y,v)_V = (y^* - (1+\tilde{\gamma})\sigma^*, v)_0 + (\hat{f},v)_0 + \langle L, v\rangle, \tag{4.18a}$$
$$\dot{\sigma} + (1+\tilde{\gamma})\sigma = y^* + \hat{f}, \tag{4.18b}$$
$$y(0) = \breve{u} + \breve{\sigma}, \tag{4.18c}$$
$$\sigma(0) = \breve{\sigma}. \tag{4.18d}$$

(4.18a),(4.18c) has a unique solution $y \in \mathcal{W}(0,T;V,V')$ by Theorem 2.19, (which is [Wlo87, Theorem 26.1] or [RR04, Theorem 11.3]). (4.18b),(4.18d) is an initial value problem with an ODE and thus has a unique solution $\sigma \in C^1([0,T];L_2(\Omega))$ by Lemma 2.18. Hence, by (2.23) the image $(y,\sigma) \in C([0,T];L_2(\Omega)^2)$ is well defined. Note also that any (y,σ) is a fixed point of A if and only if it is a solution to (4.17).

To show that A is a contraction, choose any two elements $(y_1^*, \sigma_1^*), (y_2^*, \sigma_2^*) \in C([0,T]; L_2(\Omega)^2)$ and let their images be $(y_1, \sigma_1) = A(y_1^*, \sigma_1^*)$ and $(y_2, \sigma_2) = A(y_2^*, \sigma_2^*)$. Define the differences

$$\begin{aligned} z &:= y_1 - y_2, \qquad p := \sigma_1 - \sigma_2, \\ z^* &:= y_1^* - y_2^*, \qquad p^* := \sigma_1^* - \sigma_2^*. \end{aligned}$$

In order to get the contraction property of A, we want to show that

$$\|(z,p)\| \le \frac{1}{2}\|(z^*,p^*)\|$$

holds for some norm $\|\cdot\|$. By subtracting, we get from (4.18a):

$$(\dot{z}, v)_0 + (z, v)_V = (z^* - (1+\tilde{\gamma})p^*, v)_0, \tag{4.19a}$$

and from (4.18b) we get, also taking the inner product with any $v \in V$:

$$(\dot{p}, v)_0 + ((1+\tilde{\gamma})p, v)_0 = (z^*, v)_0. \tag{4.19b}$$

We proceed in a similar way to the proof of Proposition 4.6. Choose $v = z$ in (4.19a) and $v = p$ in (4.19b) and add them:

$$\frac{1}{2}\frac{d}{dt}\|z\|_0^2 + \frac{1}{2}\frac{d}{dt}\|p\|_0^2 + \|z\|_V^2 + ((1+\tilde{\gamma})p, p)_0 = (z^* - (1+\tilde{\gamma})p^*, z)_0 + (z^*, p)_0.$$

Thus, using the Cauchy-Schwarz inequality and the bounds on $\tilde{\gamma}$,

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}\|z\|_0^2 + \frac{1}{2}\frac{d}{dt}\|p\|_0^2 + \|z\|_V^2 + (1+\check{\gamma})\|p\|_0^2 \\ \le \Big(\|z^*\|_0 + (1+\hat{\gamma})\|p^*\|_0\Big) \cdot \|z\|_0 + \|z^*\|_0 \cdot \|p\|_0. \end{aligned}$$

Multiplying by 2, using (2.12) to change the norm on $\|z\|_0$ and then applying Young's inequality (A.1), we get

$$\begin{aligned} \frac{d}{dt}\|z\|_0^2 + \frac{d}{dt}\|p\|_0^2 + 2\|z\|_V^2 + 2(1+\check{\gamma})\|p\|_0^2 \\ \le \frac{C_V^2}{2}\Big(\|z^*\|_0 + (1+\hat{\gamma})\|p^*\|_0\Big)^2 + 2\|z\|_V^2 + \|z^*\|_0^2 + \|p\|_0^2. \end{aligned}$$

The non-starred terms on the right-hand side are subtracted and after omitting the remaining $(1+2\check{\gamma})\|p\|_0^2$ on the left-hand side, the inequality becomes

$$\frac{d}{dt}\|z\|_0^2 + \frac{d}{dt}\|p\|_0^2 \le (C_V^2 + 1) \cdot \|z^*\|_0^2 + C_V^2(1+\hat{\gamma})^2 \cdot \|p^*\|_0^2.$$

Integrate over $(0,t)$ and use the initial conditions $z(0) = p(0) = 0$,

$$\|z(t)\|_0^2 + \|p(t)\|_0^2 \le c_3 \int_0^t \Big(\|z^*(s)\|_0^2 + \|p^*(s)\|_0^2\Big) ds,$$

where $c_3 = (C_V^2 + 1)(1+\hat{\gamma})^2$ can be used as a bound for both constants in the previous inequality. Recalling the definition (2.7) of the norm of $L_2(\Omega)^2$, this may be rewritten as

$$\|(z(t), p(t))\|_{L_2(\Omega)^2}^2 \le c_3 \int_0^t \|(z^*(s), p^*(s))\|_{L_2(\Omega)^2}^2 \, ds.$$

Now use weighted norms and Lemma 4.9:

$$\|(z,p)\|_{C([0,T];L_2(\Omega)^2),\beta} \leq \frac{1}{2}\|(z^*,p^*)\|_{C([0,T];L_2(\Omega)^2),\beta},$$

where $\beta = 2c_3 = 2(C_V^2+1)(1+\hat{\gamma})^2$. Thus A is a contraction on the space $C([0,T];L_2(\Omega)^2)$ when using the weighted norm. Note that $C([0,T];L_2(\Omega)^2)$ is a Banach space since $L_2(\Omega)^2$ is one and because of Lemma 2.12. Therefore, by Banach's contraction mapping theorem (Theorem A.5), A has a unique fixed point $(y,\sigma) \in C([0,T];L_2(\Omega)^2)$. The definition of A also implies that $y \in \mathcal{W}(0,T;V,V')$ and $\sigma \in C^1([0,T];L_2(\Omega))$, so that (y,σ) is the unique solution of (4.17) and $(u,\sigma) = (y-\sigma,\sigma)$ is the unique solution of the linearised version of Problem 4.3. □

4.4 Unique solution in 1D

This section gives several small results that will lead towards Proposition 4.18, which shows that the (nonlinear) problem has a unique solution on 1-dimensional domains.

The proof goes in several steps, in general following the easier example of the proof of Proposition 4.10:

1. Setting up an iteration such that any fixed point of this iteration is a solution of the problem.
2. Establishing a bound for the difference of iterates: Lemma 4.14.
3. Using this bound and weighted norms to show that the iteration is a contraction under certain conditions.
4. Using a variant of Banach's fixed point theorem (Lemma 4.17) to show that the contraction has a unique fixed point, which is the unique solution to the problem.

4.4.1 An iteration for the nonlinear problem

Lemma 4.12 (An iteration for the nonlinear problem).
Suppose that $\breve{u},\breve{\sigma} \in L_2(\Omega)$ *and* $L \in L_2(0,T;V')$ *and* $\hat{f} \in C([0,T];L_2(\Omega))$. *For any* $(u^*,\sigma^*) \in C([0,T];L_2(\Omega)^2)$ *the system of equations*

$$\langle \dot{u}, v\rangle + (u+\sigma, v)_V = \langle L, v\rangle \qquad \forall v \in V, \tag{4.20a}$$

$$\dot{\sigma} + \gamma(u^*)\sigma = u + \hat{f}, \tag{4.20b}$$

$$u(0) = \breve{u}, \tag{4.20c}$$

$$\sigma(0) = \breve{\sigma}, \tag{4.20d}$$

has a unique solution $(u,\sigma) \in C([0,T];L_2(\Omega)^2)$ *with* $u+\sigma \in \mathcal{W}(0,T;V,V')$ *and* $\sigma \in C^1([0,T];L_2(\Omega))$. *Thus,* (4.20) *defines a mapping*

$$A : C([0,T];L_2(\Omega)^2) \to C([0,T];L_2(\Omega)^2)$$

through $A(u^*,\sigma^*) := (u,\sigma)$.

Moreover, any $(u,\sigma) \in C([0,T];L_2(\Omega)^2)$ *is a solution of Problem 4.3 if and only if it is a fixed point of A.*

Remark 4.13.
Note that the equations (4.20) do not contain σ^*. Thus, the image $A(u^*,\sigma^*) = (u,\sigma)$ is actually independent of σ^*. The reason for keeping σ^* in the definition of A is that it gives a mapping which can be iterated, so that a fixed point theorem can be applied.

Proof of Lemma 4.12. Observe that since $u^* \in C([0,T]; L_2(\Omega))$, we get that $\gamma(u^*) \in C([0,T]; L_\infty(\Omega))$. Hence, equations (4.20) have a unique solution by Proposition 4.10. Finally, for a fixed point of A, we would have $u = u^*$ and then (4.20) would become Problem 4.3. □

The next step after having defined the iteration A, is to take two pre-images and look at their images under A to see whether the distance of the images is smaller by a factor < 1 than the distance of the pre-images. We will see that at least one of the pre-images needs to be bounded.

Lemma 4.14 (Bound for difference of iterates).
Let $M > 0$ be a constant and let A be the iteration from Lemma 4.12. Then there is a constant c_M, depending on M, such that

$$\|u_1(t) - u_2(t)\|_0^2 + \|\sigma_1(t) - \sigma_2(t)\|_0^2 \leq c_M \int_0^t \|u_1^*(s) - u_2^*(s)\|_0^2 \, ds, \tag{4.21}$$

for any pre-images $(u_1^*, \sigma_1^*), (u_2^*, \sigma_2^*) \in C([0,T]; L_2(\Omega)^2)$, with images $(u_1, \sigma_1) := A(u_1^*, \sigma_1^*)$ and $(u_2, \sigma_2) := A(u_2^*, \sigma_2^*)$, provided that $\sigma_2^* \in C([0,T]; L_\infty(\Omega))$ with

$$\|\sigma_2^*\|_{C([0,T];L_\infty(\Omega))} \leq M. \tag{4.22}$$

The constant can be chosen as $c_M = 3(C'_\gamma M)^2 \exp(T(48 + 3\hat{\gamma}^2))$.

Proof. Define the differences

$$\begin{aligned} z &:= u_1 - u_2, \qquad p := \sigma_1 - \sigma_2, \\ z^* &:= u_1^* - u_2^*, \qquad p^* := \sigma_1^* - \sigma_2^*. \end{aligned}$$

We also define $\gamma_1 := \gamma(u_1^*)$ and $\gamma_2 := \gamma(u_2^*)$. Then

$$\gamma_1\sigma_1 - \gamma_2\sigma_2 = \gamma_1 p + (\gamma_1 - \gamma_2)\sigma_2. \tag{4.23}$$

To start, take the difference of (4.20a) for the two images and obtain

$$(\dot{z}, v)_0 + (z + p, v)_V = 0 \quad \forall v \in V. \tag{4.24}$$

Similarly, taking the difference of (4.20b) for the two images gives

$$(\dot{p}, v)_0 + (\gamma_1\sigma_1 - \gamma_2\sigma_2, v)_0 = (z, v)_0 \quad \forall v \in V.$$

Using (4.23) and rearranging yields

$$(\dot{p}, v)_0 + (\gamma_1 p, v)_0 = (z, v)_0 - ((\gamma_1 - \gamma_2)\sigma_2, v)_0 \quad \forall v \in V. \tag{4.25}$$

Adding (4.24) and (4.25) and choosing $v = z + p$, then subtracting $(\gamma_1 p, z)_0$ on both sides gives

$$\frac{1}{2}\frac{d}{dt}\|z + p\|_0^2 + \|z + p\|_V^2 + (\gamma_1 p, p)_0 = (z, z + p)_0 - (\gamma_1 p, z)_0 - ((\gamma_1 - \gamma_2)\sigma_2, z + p)_0.$$

Now, choosing $v = p$ in (4.25) gives

$$\frac{1}{2}\frac{d}{dt}\|p\|_0^2 + (\gamma_1 p, p)_0 = (z, p)_0 - ((\gamma_1 - \gamma_2)\sigma_2, p)_0.$$

We add these last two equations and obtain

$$\frac{1}{2}\frac{d}{dt}\|z+p\|_0^2+\frac{1}{2}\frac{d}{dt}\|p\|_0^2+\|z+p\|_V^2+2(\gamma_1 p,p)_0 \\ =(z,z+2p)_0-(\gamma_1 p,z)_0-((\gamma_1-\gamma_2)\sigma_2,z+2p)_0.$$

Multiplying by 2, then leaving out the terms $\|z+p\|_V^2 \geq 0$ and $(\gamma_1 p,p)_0 \geq 0$ on the left-hand side, and finally using the Cauchy-Schwarz inequality yields

$$\frac{d}{dt}(\|z+p\|_0^2+\|p\|_0^2) \leq 2\|z\|_0\cdot\|z+2p\|_0+2\hat{\gamma}\|p\|_0\cdot\|z\|_0+2\|(\gamma_1-\gamma_2)\sigma_2\|_0\cdot\|z+2p\|_0.$$

Young's inequality (A.1) now gives

$$\frac{d}{dt}(\|z+p\|_0^2+\|p\|_0^2) \\ \leq \|z\|_0^2+\|z+2p\|_0^2\ +\hat{\gamma}^2\|p\|_0^2+\|z\|_0^2\ +\|(\gamma_1-\gamma_2)\sigma_2\|_0^2+\|z+2p\|_0^2.$$

We expand the norms $\|z+2p\|_0^2$ by using (4.6):

$$\frac{d}{dt}(\|z+p\|_0^2+\|p\|_0^2) \leq 6\|z\|_0^2+(16+\hat{\gamma}^2)\|p\|_0^2+\|(\gamma_1-\gamma_2)\sigma_2\|_0^2. \tag{4.26}$$

In this inequality, the term $\|(\gamma_1-\gamma_2)\sigma_2\|_0$ can be bounded as follows, using Lemma 3.9 and the assumption (4.22),

$$\|(\gamma_1-\gamma_2)\sigma_2\|_0 \leq \|\sigma_2\|_{L_\infty(\Omega)}\cdot\|\gamma_1-\gamma_2\|_0 \leq C'_\gamma M\|z^*\|_0. \tag{4.27}$$

If we use this in (4.26), then integrate over $(0,t)$, we obtain

$$\|z(t)+p(t)\|_0^2+\|p(t)\|_0^2 \\ \leq 6\int_0^t\|z(s)\|_0^2\,ds+(16+\hat{\gamma}^2)\int_0^t\|p(s)\|_0^2\,ds+(C'_\gamma M)^2\int_0^t\|z^*(s)\|_0^2\,ds.$$

Here the initial conditions (4.20c) and (4.20d) have been used, they imply that $z(0)=p(0)=0$. Next, applying the left part of (4.7) to the left-hand side, then multiplying by 3 gives

$$\|z(t)\|_0^2+\|p(t)\|_0^2 \\ \leq 18\int_0^t\|z(s)\|_0^2\,ds+(48+3\hat{\gamma}^2)\int_0^t\|p(s)\|_0^2\,ds+3(C'_\gamma M)^2\int_0^t\|z^*(s)\|_0^2\,ds.$$

Now Grönwall's lemma, Lemma A.2 can be used:

$$\|z(t)\|_0^2+\|p(t)\|_0^2 \leq 3(C'_\gamma M)^2\exp(T(48+3\hat{\gamma}^2))\int_0^t\|z^*(s)\|_0^2\,ds.$$

This is (4.21) and we get the expression for the constant c_M. □

Remark 4.15.
Note that (4.27) is the key estimate, where the L_∞-norm of σ appears for the first time. This will eventually lead to the requirement of a one-dimensional domain. It is possible that with a different form of γ, maybe in the form of an operator $\gamma: L_2(\Omega)\to L_\infty(\Omega)$, (4.27) can be replaced by a better estimate.

If we want to apply the previous lemma, we need to find a way to get such a constant M that bounds $\|\sigma\|_{C([0,T];L_\infty(\Omega))}$.

Lemma 4.16 ($\|\sigma\|_{L_\infty(\Omega)}$-bound of iterate in 1D).
Suppose that $d = 1$ and that $\breve{\sigma} \in L_\infty(\Omega)$ and $\hat{f} \in L_2(0,T;L_\infty(\Omega))$. If (u,σ) is an image of the iteration A from Lemma 4.12, then $\sigma \in C([0,T];L_\infty(\Omega))$ with

$$\|\sigma\|_{C([0,T];L_\infty(\Omega))} \le C(\|\breve{u}\|_0 + \|\breve{\sigma}\|_{L_\infty(\Omega)} + \|L\|_{L_2(0,T;V')} + \|\hat{f}\|_{L_2(0,T;L_\infty(\Omega))}). \quad (4.28)$$

Proof. First observe that by rewriting (4.20b), we obtain

$$\dot{\sigma} = -(\gamma(u^*)+1)\sigma + (u+\sigma) + \hat{f}.$$

Integrating this over $(0,t)$ and using the initial condition (4.20d) gives

$$\sigma(t) - \breve{\sigma} = \int_0^t \Big(-(\gamma(u^*(s))+1)\sigma(s) + (u(s)+\sigma(s)) + \hat{f}(s)\Big)\, ds.$$

Thus, taking absolute values,

$$|\sigma(t)| \le |\breve{\sigma}| + (\hat{\gamma}+1)\int_0^t |\sigma(s)|\, ds + \int_0^t |u(s)+\sigma(s)|\, ds + \int_0^t |\hat{f}(s)|\, ds.$$

Hence

$$\begin{aligned}\|\sigma(t)\|_{L_\infty(\Omega)} \le \|\breve{\sigma}\|_{L_\infty(\Omega)} + (\hat{\gamma}+1)\int_0^t \|\sigma(s)\|_{L_\infty(\Omega)}\, ds \\ + \int_0^t \|u(s)+\sigma(s)\|_{L_\infty(\Omega)}\, ds + \int_0^t \|\hat{f}(s)\|_{L_\infty(\Omega)}\, ds.\end{aligned}$$

Apply Grönwall's lemma A.2 to obtain

$$\|\sigma(t)\|_{L_\infty(\Omega)} \le c_4\Big(\|\breve{\sigma}\|_{L_\infty(\Omega)} + \int_0^t \|u(s)+\sigma(s)\|_{L_\infty(\Omega)}\, ds + \int_0^t \|\hat{f}(s)\|_{L_\infty(\Omega)}\, ds\Big), \quad (4.29)$$

where $c_4 = \exp((\hat{\gamma}+1)T)$. Using Jensen's inequality A.1 gives

$$\left(\int_0^t \|u(s)+\sigma(s)\|_{L_\infty(\Omega)}\, ds\right)^2 \le T\int_0^t \|u(s)+\sigma(s)\|^2_{L_\infty(\Omega)}\, ds.$$

Now we come to the point, where we have to restrict our attention to a 1-dimensional domain: The Sobolev embedding (Theorem 2.4) implies, since $d = 1$, and when combined with (2.11), that there is a constant $C_{\infty,V}$ such that $\|v\|_{L_\infty(\Omega)} \le C_{\infty,V}\|v\|_V$ for $v \in V$. Hence

$$\int_0^t \|u(s)+\sigma(s)\|^2_{L_\infty(\Omega)}\, ds \le C^2_{\infty,V}\int_0^t \|u(s)+\sigma(s)\|^2_V\, ds.$$

For $d = 2$ or $d = 3$, this Sobolev embedding does not hold anymore. To continue, use the last two inequalities in (4.29):

$$\|\sigma(t)\|_{L_\infty(\Omega)} \le c_4\Big(\|\breve{\sigma}\|_{L_\infty(\Omega)} + T^{1/2}C_{\infty,V}\|u+\sigma\|_{L_2(0,T;V)} + \|\hat{f}\|_{L_2(0,T;L_\infty(\Omega))}\Big). \quad (4.30)$$

The remaining difficulty is to obtain a bound on $\|u+\sigma\|_{L_2(0,T;V)}$. Looking again at the proof of Proposition 4.6, we see that all steps of the proof are still valid, when we apply them to (4.20) instead of Problem 4.3. Having $\gamma(u^*)$ instead of $\gamma(u)$ in the

equations does not change anything in the proof. Thus, (4.15) also holds for (4.20) and this implies

$$\|u+\sigma\|_{L_2(0,T;V)} \leq c_1^{1/2}(\|\breve{u}\|_0 + \|\breve{\sigma}\|_0 + \|L\|_{L_2(0,T;V')}) + c_2^{1/2}\|\hat{f}\|_{L_2(0,T;L_2(\Omega))}.$$

We insert this into (4.30) and since there is a constant $C_{0,\infty}$ such that

$$\|\breve{\sigma}\|_0 \leq C_{0,\infty}\|\breve{\sigma}\|_{L_\infty(\Omega)}, \quad \|\hat{f}\|_{L_2(0,T;L_2(\Omega))} \leq C_{0,\infty}\|\hat{f}\|_{L_2(0,T;L_\infty(\Omega))},$$

we finally obtain:

$$\begin{aligned}\|\sigma\|_{C([0,T];L_\infty(\Omega))} \leq\ & c_4(c_1T)^{1/2}C_{\infty,V}\cdot\|\breve{u}\|_0 \\ & + c_4(1+(c_1T)^{1/2}C_{0,\infty}C_{\infty,V})\cdot\|\breve{\sigma}\|_{L_\infty(\Omega)} \\ & + c_4(c_1T)^{1/2}C_{\infty,V}\cdot\|L\|_{L_2(0,T;V')} \\ & + c_4(1+(c_2T)^{1/2}C_{0,\infty}C_{\infty,V})\cdot\|\hat{f}\|_{L_2(0,T;L_\infty(\Omega))}.\end{aligned} \tag{4.31}$$

This concludes the proof, giving a specific expression for the constant in (4.28), if we insert the value of c_4 from this proof and the values of c_1 and c_2 from the proof of Proposition 4.6. □

4.4.2 The fixed point of the iteration

For showing the uniqueness of the solutions in 1D, we would like to proceed in a similar way to the proof of Proposition 4.10. This works only up to a certain point. We can combine the estimate Lemma 4.14 with the weighted norm trick (Lemma 4.9) and see that the iteration A is a contraction, see later in the proof of Proposition 4.18. But the difficulty is that $\|\sigma\|_{C([0,T];L_\infty(\Omega))}$ needs to be bounded. Thus, A is not a contraction on all of $C([0,T];L_2(\Omega)^2)$ but only on a subset for which $\|\sigma\|_{C([0,T];L_\infty(\Omega))}$ is bounded. This subset might not be a closed subset, so that the standard version of Banach's fixed point theorem (Theorem A.5) can not be applied.

Instead we now give a lemma, that is very similar to Banach's fixed point theorem. The main difference is that the subset Y does not have to be a closed subset anymore. Note that in the contraction condition below, one of the pre-images must be in the subset Y, while the other one can be from the whole space X. This corresponds to the conditions under which (4.21) holds: $\|\sigma_2^*\|_{C([0,T];L_\infty(\Omega))}$ must be bounded, but $\|\sigma_1^*\|_{C([0,T];L_\infty(\Omega))}$ need not be bounded.

Lemma 4.17 (Banach fixed point theorem for subsets).
Let $(X,\|\cdot\|_X)$ be a Banach space and $Y \subseteq X$ a subset. Suppose $A: X \to X$ is a mapping such that

- *its image is contained in Y: $A(X) \subseteq Y$,*
- *A is a contraction, if at least one pre-image is in Y. By that we mean that there is $0 \leq q < 1$ such that*

$$\|A(x)-A(y)\|_X \leq q\cdot\|x-y\|_X \quad \forall\, x \in X, y \in Y.$$

Then A has a fixed point $z \in Y$, which is unique in X.

Proof. (a) Uniqueness: assume $x_1, x_2 \in X$ are fixed points, i.e. $A(x_1)=x_1$, $A(x_2)=x_2$. Then $x_1, x_2 \in A(X) \subseteq Y$. Thus $\|x_1-x_2\|_X = \|A(x_1)-A(x_2)\|_X \leq q\|x_1-x_2\|_X$, and hence $\|x_1-x_2\|=0$.

(b) Existence: Choose any $y_0 \in Y$ and create the sequence $(y_n)_{n\in\mathbb{N}}$ using the iteration $y_{n+1} = A(y_n)$. Since A is a contraction on the set Y, we get that y_n is a Cauchy sequence with respect to $\|\cdot\|_X$, just like in the proof of the standard Banach fixed point theorem, see for example [OD96, Theorem 4.10.1]. Because of the completeness of X, this implies the convergence of (y_n) to an element $z \in X$. It remains to show that z is a fixed point: $z = A(z)$. For any $n \in \mathbb{N}$ we have

$$\begin{aligned}\|z - A(z)\|_X &= \|(z - y_n) + (y_n - y_{n+1}) + (y_{n+1} - A(z))\|_X \\ &\leq \|z - y_n\|_X + \|y_n - y_{n+1}\|_X + \|y_{n+1} - A(z)\|_X \\ &\leq \|z - y_n\|_X + \|y_n - y_{n+1}\|_X + q\|y_n - z\|_X,\end{aligned}$$

where we have used that $y_n \in Y$ in order to apply the contraction property. Now, because y_n converges to z and the Cauchy sequence property of y_n, the terms on the right-hand side can be made arbitrarily small by choosing sufficiently large values for n. Hence $\|z - A(z)\|_X = 0$ and $z = A(z)$. □

With the modified fixed point theorem and the lemmas from the previous subsection we have now collected all the results that are needed to prove existence and uniqueness in 1D:

Proposition 4.18 (Unique solution in 1D).
Suppose that $d = 1$ and that $\breve{\sigma} \in L_\infty(\Omega)$ and $\hat{f} \in L_2(0,T;L_\infty(\Omega))$ in addition to the assumptions of Problem 4.3. Then Problem 4.3 has a unique solution

$$(u, \sigma) \in C([0,T]; L_2(\Omega)^2),$$

with $u + \sigma \in \mathcal{W}(0,T;V,V')$ and $\sigma \in C^1([0,T];L_2(\Omega))$.

Proof. Let M be the right-hand side of (4.28). For an explicit expression of the value of M, refer to the right-hand side of (4.31). We would like to apply Lemma 4.17 to the iteration A, defined in Lemma 4.12. We define the space X and a subset Y by

$$\begin{aligned}X &:= C([0,T]; L_2(\Omega)^2), \\ Y &:= \{(u,\sigma) \in X \ : \ \|\sigma\|_{C([0,T];L_\infty(\Omega))} \leq M\}.\end{aligned}$$

From Lemma 4.16, we see that $A(X) \subseteq Y$. Denoting the pairs (u, σ) by bold letters like $\boldsymbol{a}$, $\boldsymbol{b}$ now, Lemma 4.14 implies,

$$\|A(\boldsymbol{a})(t) - A(\boldsymbol{b})(t)\|^2_{L_2(\Omega)^2} \leq c_M \int_0^t \|\boldsymbol{a}(s) - \boldsymbol{b}(s)\|^2_{L_2(\Omega)^2}\, ds \qquad \forall \boldsymbol{a} \in X, \boldsymbol{b} \in Y. \quad (4.32)$$

With the help of Lemma 4.9 and weighted norms, this gives a contraction property:

$$\|A(\boldsymbol{a}) - A(\boldsymbol{b})\|_{C([0,T];L_2(\Omega)^2),\beta} \leq \frac{1}{2}\|\boldsymbol{a} - \boldsymbol{b}\|_{C([0,T];L_2(\Omega)^2),\beta} \qquad \forall \boldsymbol{a} \in X, \boldsymbol{b} \in Y. \quad (4.33)$$

where $\beta := 2c_M$. Thus, if we use the weighted norm on the space X instead of the standard norm,

$$\|\boldsymbol{a}\|_X := \|\boldsymbol{a}\|_{C([0,T];L_2(\Omega)^2),\beta} \quad \text{for } \boldsymbol{a} \in X,$$

then we can apply Lemma 4.17. Therefore, A has a unique fixed point $\boldsymbol{a} \in Y$. Using Lemma 4.12 now concludes the proof. □

4.4.3 What about $d = 2$ and $d = 3$?

Since Proposition 4.18 is restricted to $d = 1$, the obvious question to ask now is what can be done about the cases $d = 2$ and $d = 3$. The proof of Proposition 4.18 relies on the Sobolev embedding of the space $H^1(\Omega)$ into the space $L_\infty(\Omega)$, which does not hold for $d \geq 2$.

In order to extend the result to higher dimensions we need a higher regularity of the solution. Thus, for obtaining a similar result to Proposition 4.18 for $d \geq 2$, one of the following approaches could help:

- Leave the Hilbert space setting of $H^1(\Omega)$ and work in $W_p^1(\Omega)$ for $p > d$. The Sobolev embedding would give us

  $$\|v\|_{L_\infty(\Omega)} \leq C\|v\|_{W_p^1(\Omega)} \quad \forall\, v \in W_p^1(\Omega).$$

  The difficulty with this approach is that the calculations can no longer use the inner products.

- Stay in the Hilbert space setting but of a higher order: $H^2(\Omega)$. We could use the Sobolev embedding:

  $$\|v\|_{L_\infty(\Omega)} \leq C\|v\|_{H^2(\Omega)} \quad \forall\, v \in H^2(\Omega),$$

  since $d \leq 3$.

  Thus to find a replacement for Proposition 4.18 in the case $d \geq 2$, a bound for $\|\sigma\|_{H^2(\Omega)}$ is required.

  In general the Hilbert space setting is suitable for calculating such bounds.

  Instead of choosing $v = u + \sigma$ in the sum of (4.20a) and (4.20b) one would choose $v = -\Delta(u + \sigma)$ and instead of taking the L_2 inner product of (4.20b) with $v = \sigma$ one would choose $v = -\Delta\sigma$.

  If $\mathcal{ER}^1$ holds, then Green's formula implies

  $$(u, v)_V = (-\Delta u, v)_0 \qquad \forall\, u \in V \cap H^2(\Omega),\ v \in V,$$

  and also

  $$(\dot{v}, -\Delta v)_0 = \frac{1}{2}\frac{d}{dt}\|v\|_V^2 \qquad \forall\, v \in V \cap H^2(\Omega).$$

  With the above choices of v and using these equalities, we obtain

  $$\frac{1}{2}\frac{d}{dt}\|u + \sigma\|_V^2 + \|\Delta(u + \sigma)\|_0^2 + (\gamma(u^*)\sigma, u + \sigma)_V = (L + \hat{f} + u, u + \sigma)_V,$$
  $$\frac{1}{2}\frac{d}{dt}\|\sigma\|_V^2 + (\gamma(u^*)\sigma, \sigma)_V = (\hat{f} + u, \sigma)_V.$$

  Two difficulties remain: Firstly, it is necessary to take the space derivative of the nonlinear term $\gamma(u^*)\sigma$,

  $$\nabla\gamma(u^*)\sigma = \gamma'(u^*)\sigma\nabla u^* + \gamma(u^*)\nabla\sigma,$$

  which gives extra terms which are hard to deal with. Secondly, we would obtain bounds for $\|\Delta(u + \sigma)\|_0$. But is this enough to obtain a bound for $\|\sigma\|_{H^2(\Omega)}$?

To conclude, we see that both approaches entail difficulties and the question of whether our variational formulation can have a unique solution in 2D and 3D remains open.

Chapter 5

Discretisation and Numerical Scheme

In this chapter the discrete scheme for the approximate solution of the Cohen model is presented.

The space discretisation uses a family of meshes $\mathcal{E}_h$ as described by Assumption 2.21, and the associated finite element spaces V^h which were defined in Definition 2.22.

The notation, definitions and some background for the time discretisation is given in Section 5.1.

The discrete scheme, Problem 5.4, is stated in Section 5.2. Five methods for the discretisation of the nonlinear term are considered.

In Section 5.3 the scheme is reformulated as a matrix equation. For one of the five methods, the matrix equation can be simplified, so that the matrices involved become smaller.

The question of whether the discrete scheme really defines a unique discrete solution is examined in Section 5.4.

Finally Section 5.5 contains some results on the boundedness of the discrete solution.

5.1 Time discretisation

For the time discretisation the time interval $[0,T]$ is divided into N subintervals with equidistant endpoints t_n such that $0 = t_0 < t_1 < \cdots < t_N = T$. Thus, the interval length $t_n - t_{n-1}$ is constant and will be denoted by k:

$$k := \frac{T}{N} = t_n - t_{n-1} \quad \forall n \in \{1, \ldots, N\}. \tag{5.1}$$

5.1.1 Notation

To simplify notation, define for any time-dependent function $v : [0,T] \to V'$:

$$v_n := v(t_n). \tag{5.2}$$

Using these, define also

$$\begin{aligned}
\partial_t v_n &:= \frac{1}{k}(v_n - v_{n-1}),\\
\partial_t^2 v_n &:= \frac{1}{k^2}(v_n - 2v_{n-1} + v_{n-2}),\\
\bar{v}_n &:= \frac{1}{2}(v_n + v_{n-1}),\\
\Delta_n v &:= \bar{\dot{v}}_n - \partial_t v_n = \frac{\dot{v}(t_n) + \dot{v}(t_{n-1})}{2} - \frac{v(t_n) - v(t_{n-1})}{k}.
\end{aligned} \tag{5.3}$$

Observe that $\partial_t v_n$ is an approximation for the derivative $\dot{v}$ at $t = (t_n + t_{n-1})/2$, and $\partial_t^2 v_n$ is an approximation for the second time derivative $\ddot{v}$ at $t = t_{n-1}$. Since we will need it frequently later, we note that, using (4.6),

$$\|\bar{v}_n\|^2 \le \frac{1}{2}\|v_n\|^2 + \frac{1}{2}\|v_{n-1}\|^2, \tag{5.4}$$

which, in turn, implies

$$\sum_{n=1}^{m}\|\bar{v}_n\|^2 \le \frac{1}{2}\|v_0\|^2 + \sum_{n=1}^{m-1}\|v_n\|^2 + \frac{1}{2}\|v_m\|^2 \le \sum_{n=0}^{m}\|v_n\|^2. \tag{5.5}$$

Another direct consequence of the above definitions is

$$(\partial_t v_n, \bar{v}_n)_0 = \frac{1}{2k}\left(\|v_n\|_0^2 - \|v_{n-1}\|_0^2\right), \tag{5.6}$$

which can be viewed as a discrete version of (2.21).

5.1.2 Taylor estimates

Lemma 5.1 (Taylor estimates).
Let X be a Banach space. If v has the regularity, which is indicated by the norm on the right-hand side, then

$$\|\partial_t v_n\|_X \le \|\dot{v}\|_{L_\infty(t_{n-1},t_n;X)}, \tag{5.7}$$
$$\|\partial_t v_n\|_0 \le k^{-1/2}\|\dot{v}\|_{L_2(t_{n-1},t_n;L_2(\Omega))}, \tag{5.8}$$
$$\|\partial_t^2 v_n\|_X \le \|\ddot{v}\|_{L_\infty(t_{n-2},t_n;X)}, \tag{5.9}$$
$$\|\partial_t^2 v_n\|_0 \le k^{-1/2}\|\ddot{v}\|_{L_2(t_{n-2},t_n;L_2(\Omega))}, \tag{5.10}$$
$$\|\Delta_n v\|_0 \le Ck^{3/2}\|\dddot{v}\|_{L_2(t_{n-1},t_n;L_2(\Omega))}. \tag{5.11}$$

Proof. For proving (5.7) we have from Taylor's theorem (or the mean value theorem)

$$v_n = v_{n-1} + k\dot{v}(\xi) \quad \text{for some } \xi \in [t_{n-1}, t_n].$$

Thus $\partial_t v_n = (v_n - v_{n-1})/k = \dot{v}(\xi)$, whence $\|\partial_t v_n\|_X = \|\dot{v}(\xi)\|_X \le \|\dot{v}\|_{L_\infty(t_{n-1},t_n;X)}$. For proving (5.9) Taylor's theorem also gives

$$\begin{aligned}
v_n &= v_{n-1} + k\dot{v}_{n-1} + \frac{k^2}{2}\ddot{v}(\xi_1) \quad \text{for some } \xi_1 \in [t_{n-1}, t_n],\\
v_{n-2} &= v_{n-1} - k\dot{v}_{n-1} + \frac{k^2}{2}\ddot{v}(\xi_2) \quad \text{for some } \xi_2 \in [t_{n-2}, t_{n-1}].
\end{aligned}$$

Thus $\partial_t^2 v_n = (v_n - 2v_{n-1} + v_{n-2})/k^2 = \frac{1}{2}(\ddot{v}(\xi_1) + \ddot{v}(\xi_2))$, whence $\|\partial_t^2 v_n\|_X \leq \frac{1}{2}(\|\ddot{v}(\xi_1)\|_X + \|\ddot{v}(\xi_2)\|_X) \leq \|\ddot{v}\|_{L_\infty(t_{n-2},t_n;X)}$. To prove (5.8) we use Taylor's theorem with integral remainder (or the fundamental theorem of calculus):

$$v_n = v_{n-1} + \int_{t_{n-1}}^{t_n} \dot{v}(t)\, dt.$$

Thus, using the Cauchy-Schwarz inequality,

$$|v_n - v_{n-1}| = \Big| \int_{t_{n-1}}^{t_n} \dot{v} dt \Big| \leq \sqrt{\int_{t_{n-1}}^{t_n} 1^2 dt} \cdot \sqrt{\int_{t_{n-1}}^{t_n} \dot{v}^2 dt} = \sqrt{k \int_{t_{n-1}}^{t_n} \dot{v}^2 dt}.$$

Squaring both sides and integrating over the domain now yields

$$\|v_n - v_{n-1}\|_0^2 \leq k \int_\Omega \int_{t_{n-1}}^{t_n} \dot{v}^2 dt\, d\Omega = k \int_{t_{n-1}}^{t_n} \|\dot{v}\|_0^2 dt = k \|\dot{v}\|^2_{L_2(t_{n-1},t_n;L_2(\Omega))}.$$

Taking the square root and dividing by k gives (5.8). In order to prove (5.10), we use again Taylor's theorem with integral remainder terms:

$$v_n = v_{n-1} + k\dot{v}_{n-1} + \int_{t_{n-1}}^{t_n} (t_n - t)\, \ddot{v}(t)\, dt,$$
$$v_{n-2} = v_{n-1} - k\dot{v}_{n-1} + \int_{t_{n-2}}^{t_{n-1}} (t - t_{n-2})\, \ddot{v}(t)\, dt.$$

Therefore $k^2 \partial_t^2 v_n = v_n - 2v_{n-1} + v_{n-2} = \int_{t_{n-1}}^{t_n} (t_n - t)\, \ddot{v}\, dt + \int_{t_{n-2}}^{t_{n-1}} (t - t_{n-2})\, \ddot{v}\, dt$, which gives, using the Cauchy-Schwarz inequality,

$$\begin{aligned} k^4 |\partial_t^2 v_n|^2 &\leq 2\Big(\int_{t_{n-1}}^{t_n} (t_n - t)\, \ddot{v}\, dt \Big)^2 + 2\Big(\int_{t_{n-2}}^{t_{n-1}} (t - t_{n-2})\, \ddot{v}\, dt \Big)^2 \\ &\leq 2 \int_{t_{n-1}}^{t_n} (t_n - t)^2\, dt \cdot \int_{t_{n-1}}^{t_n} \ddot{v}^2\, dt + 2 \int_{t_{n-2}}^{t_{n-1}} (t - t_{n-2})^2\, dt \cdot \int_{t_{n-2}}^{t_{n-1}} \ddot{v}^2\, dt. \end{aligned}$$

Evaluating the integrals now gives

$$k^4 |\partial_t^2 v_n|^2 \leq \frac{2k^3}{3} \int_{t_{n-2}}^{t_n} \ddot{v}^2\, dt.$$

Dividing by k^4 and integrating over the domain gives

$$\|\partial_t^2 v_n\|_0^2 \leq \frac{2}{3k} \int_\Omega \int_{t_{n-2}}^{t_n} \ddot{v}^2\, dt\, d\Omega = \frac{2}{3k} \int_{t_{n-2}}^{t_n} \|\ddot{v}\|_0^2\, dt = \frac{2}{3k} \|\ddot{v}\|^2_{L_2(t_{n-2},t_n;L_2(\Omega))}.$$

By taking the square root and omitting the constant $\sqrt{2/3} < 1$, we arrive at (5.10). (5.11) is another standard estimate, see for example [RS05]. □

5.1.3 Discrete inverse Laplacian

For obtaining bounds on the discrete solution and for the error analysis, we will later need a discrete version of the inverse Laplacian, which was introduced in Subsection 2.1.4. The discrete inverse Laplacian $\mathcal{G}^h$, is defined as the discrete finite element solution of the Laplace equation:

Definition 5.2 (Discrete inverse Laplacian).
For any $w \in V'$, define $\mathcal{G}^h w \in V^h$ via,

$$(\mathcal{G}^h w, v)_V = \langle w, v\rangle \quad \forall\, v \in V^h. \tag{5.12}$$

It follows that $\mathcal{G}^h : V' \to V^h$ is linear.

Lemma 5.3 (Properties of discrete inverse Laplacian).
The discrete inverse Laplacian has the following stability property:

$$\|\mathcal{G}^h w\|_V \le \|w\|_{V'} \qquad \forall\, w \in V'. \tag{5.13}$$

Furthermore, for any $v_n, v_{n-1} \in V^h$,

$$(\partial_t v_n, \mathcal{G}^h \bar{v}_n)_0 = \frac{1}{2k}\|\mathcal{G}^h v_n\|_V^2 - \frac{1}{2k}\|\mathcal{G}^h v_{n-1}\|_V^2. \tag{5.14}$$

Proof. Using the definition of $\mathcal{G}^h$, we have

$$\|\mathcal{G}^h w\|_V^2 = (\mathcal{G}^h w, \mathcal{G}^h w)_V = \langle w, \mathcal{G}^h w\rangle \le \|w\|_{V'} \cdot \|\mathcal{G}^h w\|_V.$$

Dividing by $\|\mathcal{G}^h w\|_V$ concludes the proof of (5.13).
Next, using the definitions of $\partial_t v_n$ and of $\bar{v}_n$, the linearity and definition of $\mathcal{G}^h$, we get,

$$\begin{aligned}
(\partial_t v_n, \mathcal{G}^h \bar{v}_n)_0 &= \frac{1}{2k}(v_n - v_{n-1}, \mathcal{G}^h v_n + \mathcal{G}^h v_{n-1})_0 \\
&= \frac{1}{2k}((v_n, \mathcal{G}^h v_n)_0 + (v_n, \mathcal{G}^h v_{n-1})_0 - (v_{n-1}, \mathcal{G}^h v_n)_0 - (v_{n-1}, \mathcal{G}^h v_{n-1})_0) \\
&= \frac{1}{2k}((\mathcal{G}^h v_n, \mathcal{G}^h v_n)_V + (\mathcal{G}^h v_n, \mathcal{G}^h v_{n-1})_V \\
&\quad - (\mathcal{G}^h v_{n-1}, \mathcal{G}^h v_n)_V - (\mathcal{G}^h v_{n-1}, \mathcal{G}^h v_{n-1})_V) \\
&= \frac{1}{2k}(\|\mathcal{G}^h v_n\|_V^2 - \|\mathcal{G}^h v_{n-1}\|_V^2).
\end{aligned}$$

□

5.2 Discrete formulation

This section describes the discrete formulation for the Cohen model that will be used for this chapter and the next. This version of the formulation prescribes homogeneous Dirichlet boundary conditions for both u and σ to make the proof of some of the later theorems possible. As a basis we take the simplified variational formulation from Problem 4.3, instead of the more general Problem 4.1. The discrete formulation combines a standard Galerkin finite element method for the space discretisation with the Crank-Nicolson method for the time discretisation.

Problem 5.4 (Discrete formulation for error analysis).
For $n = 0, \dots, N$ find discrete solutions $u_n^h \in V^h$ and $\sigma_n^h \in V^h$ such that

$$(\partial_t u_n^h, v)_0 + (\bar{u}_n^h + \bar{\sigma}_n^h, v)_V = \langle \bar{L}_n, v\rangle \qquad \forall\, v \in V^h,\ n = 1, \dots, N, \tag{5.15a}$$

$$(\partial_t \sigma_n^h, v)_0 + (\mathscr{B}_n^Q(u^h, \sigma^h), v)_0 = (\bar{u}_n^h, v)_0 + (\bar{\hat{f}}_n, v)_0 \qquad \forall\, v \in V^h,\ n = 1, \dots, N, \tag{5.15b}$$

$$u_0^h = \Pi_0^h \breve{u}, \tag{5.15c}$$

$$\sigma_0^h = \Pi_0^h \breve{\sigma}. \tag{5.15d}$$

Here, the term $\mathscr{B}_n^Q(u^h, \sigma^h)$ in (5.15b) is the discrete approximation of the nonlinearity $\overline{\gamma(u)\sigma}_n = \frac{1}{2}(\gamma(u_n)\sigma_n + \gamma(u_{n-1})\sigma_{n-1})$. To deal with this term, we consider five possibilities, indexed by $Q \in \{1, 2, 3, 4, 5\}$.

Definition 5.5 ($\mathscr{B}_n^Q$ and $\mathscr{E}_n^Q$).
Firstly, for $Q \in \{1, 4\}$, $\mathscr{B}_n^Q$ is defined by

$$\mathscr{B}_n^Q := \frac{1}{2}\gamma(\mathscr{E}_n^Q u^h)\sigma_n^h + \frac{1}{2}\gamma(u_{n-1}^h)\sigma_{n-1}^h, \tag{5.16}$$

where $\mathscr{E}_n^Q$ denotes an extrapolation operator, defined as

$$\begin{aligned}\mathscr{E}_n^1 u^h &:= \begin{cases} 2u_{n-1}^h - u_{n-2}^h & \text{for } n \geq 2, \\ u_{n-1}^h & \text{for } n = 1, \end{cases} \\ \mathscr{E}_n^4 u^h &:= u_n^h.\end{aligned}$$

Secondly, for $Q \in \{2, 5\}$ we set

$$\mathscr{B}_n^Q := \gamma(\mathscr{E}_n^Q u^h)\bar{\sigma}_n^h \tag{5.17}$$

where, again, $\mathscr{E}_n^Q$ is an extrapolation operator

$$\begin{aligned}\mathscr{E}_n^2 u^h &:= \begin{cases} \frac{3}{2}u_{n-1}^h - \frac{1}{2}u_{n-2}^h & \text{for } n \geq 2, \\ u_{n-1}^h & \text{for } n = 1, \end{cases} \\ \mathscr{E}_n^5 u^h &:= \bar{u}_n^h.\end{aligned}$$

Thirdly, for the remaining method $Q = 3$, we define $\mathscr{B}_n^Q$ by

$$\mathscr{B}_n^3 := I^h \mathscr{B}_n^1, \tag{5.18}$$

where I^h is the interpolation operator defined in (2.31).

$Q = 4$ is a straightforward replacement of u_n by u_n^h and of σ_n by σ_n^h. This involves solving a nonlinear system of equations at each time step, since the new discrete solution u_n^h appears inside function γ. To get the linearised scheme $Q = 1$, linear extrapolation is used on u^h.

The idea for $Q = 2$ and $Q = 5$ was taken from [RS05]. Method $Q = 5$ is nonlinear while $Q = 2$ uses extrapolation.

Method $Q = 3$ is included in this thesis, because the resulting linear system is smaller and therefore easier to solve, see later in Lemma 5.8. From a theoretical point of view however, $Q = 3$ appears to be more difficult and we have no results. It involves a so-called *product approximation* see [CGMSS81]. See also Section 6.5.

Remark 5.6 (Why not $Q = 6$?).
Analogous to the definition in (5.18) it is possible to define a sixth method $Q = 6$, by $\mathscr{B}_n^6 := I^h \mathscr{B}_n^2$. This method would be very similar to $Q = 3$. Like $Q = 3$ it would give a smaller linear system and cause the same theoretical difficulties. Because of these similarities, we do not include this method.

5.3 Matrix formulation

In this section we give a matrix formulation of the discrete problem 5.4. Let ϕ_i, $(i = 1, \ldots, N_V)$ denote the Lagrange basis functions of the finite element space V^h and let $\boldsymbol{a}_i$, $(i = 1, \ldots, N_V)$ be their nodes such that $\phi_i(\boldsymbol{a}_j) = \delta_{ij}$. For the definition of $\boldsymbol{a}_i$ and ϕ_i and some of their properties, see Definition 2.23 and Lemma 2.24.
Define the stiffness matrix $\underline{\boldsymbol{A}}$ and the mass matrix $\underline{\boldsymbol{C}}$ as usual by

$$\begin{aligned}\underline{\boldsymbol{A}}_{ij} &:= (\phi_j, \phi_i)_V, \\ \underline{\boldsymbol{C}}_{ij} &:= (\phi_j, \phi_i)_0, \qquad i, j \in \{1, \ldots, N_V\}.\end{aligned} \tag{5.19}$$

For $Q = 3$ we will need the diagonal matrices $\underline{\boldsymbol{\Gamma}}^n$ and $\underline{\boldsymbol{\Upsilon}}^n$:

$$\begin{aligned} \underline{\boldsymbol{\Gamma}}^n_{ij} &:= \delta_{ij}\gamma(\mathscr{E}^1_n u^h(\boldsymbol{a}_i)), \\ \underline{\boldsymbol{\Upsilon}}^n_{ij} &:= \delta_{ij}\gamma(u^h_{n-1}(\boldsymbol{a}_i)), \qquad i,j \in \{1,\dots,N_V\}. \end{aligned} \tag{5.20}$$

Define also load vectors for the right-hand side of the system:

$$\begin{aligned} \boldsymbol{L}^n_i &:= (\bar{f}_n, \phi_i)_0 + (\bar{g}_n, \phi_i)_{\Gamma_N}, \\ \hat{\boldsymbol{L}}^n_i &:= (\bar{\hat{f}}_n, \phi_i)_0, \qquad i \in \{1,\dots,N_V\}. \end{aligned} \tag{5.21}$$

Lemma 5.7 (Matrix form of discretised problem).
Suppose that the discrete solution is expressed using vectors $\boldsymbol{U}^n, \boldsymbol{S}^n \in \mathbb{R}^{N_V}$ such that

$$u^h_n = \sum_{j=1}^{N_V} \boldsymbol{U}^n_j \phi_j, \qquad \sigma^h_n = \sum_{j=1}^{N_V} \boldsymbol{S}^n_j \phi_j. \tag{5.22}$$

Then the discrete formulation of Problem 5.4 becomes, for $n > 0$, the following system of $2N_V$ equations:

$$\begin{aligned} &\begin{pmatrix} \frac{1}{k}\underline{\boldsymbol{C}} + \frac{1}{2}\underline{\boldsymbol{A}} & \frac{1}{2}\underline{\boldsymbol{A}} \\ -\frac{1}{2}\underline{\boldsymbol{C}} & \frac{1}{k}\underline{\boldsymbol{C}} + \frac{1}{2}\underline{\boldsymbol{X}}^n_Q \end{pmatrix} \cdot \begin{pmatrix} \boldsymbol{U}^n \\ \boldsymbol{S}^n \end{pmatrix} \\ &\qquad = \begin{pmatrix} \frac{1}{k}\underline{\boldsymbol{C}} - \frac{1}{2}\underline{\boldsymbol{A}} & -\frac{1}{2}\underline{\boldsymbol{A}} \\ \frac{1}{2}\underline{\boldsymbol{C}} & \frac{1}{k}\underline{\boldsymbol{C}} - \frac{1}{2}\underline{\boldsymbol{Y}}^n_Q \end{pmatrix} \cdot \begin{pmatrix} \boldsymbol{U}^{n-1} \\ \boldsymbol{S}^{n-1} \end{pmatrix} + \begin{pmatrix} \boldsymbol{L}^n \\ \hat{\boldsymbol{L}}^n \end{pmatrix}, \end{aligned} \tag{5.23}$$

where the matrices $\underline{\boldsymbol{X}}^n_Q$ and $\underline{\boldsymbol{Y}}^n_Q$ depend on the method Q: For $Q \in \{1,2,4,5\}$ they are modified mass matrices,

$$\underline{\boldsymbol{X}}^{Q,n}_{ij} := \begin{cases} (\gamma(\mathscr{E}^1_n u^h)\phi_j, \phi_i)_0 & \text{for } Q = 1, \\ (\gamma(\mathscr{E}^2_n u^h)\phi_j, \phi_i)_0 & \text{for } Q = 2, \\ (\gamma(u^h_n)\phi_j, \phi_i)_0 & \text{for } Q = 4, \\ (\gamma(\bar{u}^h_n)\phi_j, \phi_i)_0 & \text{for } Q = 5, \end{cases}$$

$$\underline{\boldsymbol{Y}}^{Q,n}_{ij} := \begin{cases} (\gamma(u^h_{n-1})\phi_j, \phi_i)_0 & \text{for } Q = 1, \\ (\gamma(\mathscr{E}^2_n u^h)\phi_j, \phi_i)_0 & \text{for } Q = 2, \\ (\gamma(u^h_{n-1})\phi_j, \phi_i)_0 & \text{for } Q = 4, \\ (\gamma(\bar{u}^h_n)\phi_j, \phi_i)_0 & \text{for } Q = 5, \end{cases} \qquad i,j \in \{1,\dots,N_V\}.$$

For $Q = 3$ they can be expressed as the product of the mass matrix and a diagonal matrix defined above,

$$\underline{\boldsymbol{X}}^{3,n} := \underline{\boldsymbol{C}} \cdot \underline{\boldsymbol{\Gamma}}^n, \qquad \underline{\boldsymbol{Y}}^{3,n} := \underline{\boldsymbol{C}} \cdot \underline{\boldsymbol{\Upsilon}}^n. \tag{5.24}$$

Proof. Inserting (5.22) into (5.15a) and choosing the Lagrange basis function $v = \phi_i$ for some $i \in \{1,\dots,N_V\}$ gives, using the linearity of the inner products,

$$\begin{aligned} \frac{1}{k}\sum_{j=1}^{N_V}(\phi_j,\phi_i)_0 \cdot (\boldsymbol{U}^n_j - \boldsymbol{U}^{n-1}_j) + \frac{1}{2}\sum_{j=1}^{N_V}(\phi_j,\phi_i)_V \cdot (\boldsymbol{U}^n_j + \boldsymbol{U}^{n-1}_j + \boldsymbol{S}^n_j + \boldsymbol{S}^{n-1}_j) \\ = (\bar{f},\phi_i)_0 + (\bar{g},\phi_i)_{\Gamma_N}. \end{aligned}$$

Combining the equations for $i = 1, \ldots, N_V$ into a matrix equation and rearranging gives

$$(\tfrac{1}{k}\underline{\boldsymbol{C}} + \tfrac{1}{2}\underline{\boldsymbol{A}}) \cdot \boldsymbol{U}^n + \tfrac{1}{2}\underline{\boldsymbol{A}} \cdot \boldsymbol{S}^n = (\tfrac{1}{k}\underline{\boldsymbol{C}} - \tfrac{1}{2}\underline{\boldsymbol{A}}) \cdot \boldsymbol{U}^{n-1} - \tfrac{1}{2}\underline{\boldsymbol{A}} \cdot \boldsymbol{S}^{n-1} + \boldsymbol{L}^n. \tag{5.25}$$

This is already the first half of the combined matrix equation (5.23). To get the other half, we insert (5.22) into (5.15b) and again choose $v = \phi_i$:

$$\begin{aligned}\frac{1}{k}\sum_{j=1}^{N_V}(\phi_j, \phi_i)_0 \cdot (\boldsymbol{S}_j^n - \boldsymbol{S}_j^{n-1}) + (\mathscr{B}_n^Q, \phi_i)_0 \\ = \frac{1}{2}\sum_{j=1}^{N_V}(\phi_j, \phi_i)_0 \cdot (\boldsymbol{U}_j^n + \boldsymbol{U}_j^{n-1}) + (\bar{\hat{f}}, \phi_i)_0.\end{aligned} \tag{5.26}$$

Next look in detail at $(\mathscr{B}_n^Q, \phi_i)_0$: First, for $Q \in \{1, 2, 4, 5\}$ the definition of $\mathscr{B}_n^Q$ yields

$$\mathscr{B}_n^Q = \frac{1}{2}(G_n^Q \sigma_n^h + H_n^Q \sigma_{n-1}^h),$$

where

$$G_n^Q := \begin{cases} \gamma(\mathscr{E}_n^1 u^h) & \text{for } Q = 1, \\ \gamma(\mathscr{E}_n^2 u^h) & \text{for } Q = 2, \\ \gamma(u_n^h) & \text{for } Q = 4, \\ \gamma(\bar{u}_n^h) & \text{for } Q = 5, \end{cases} \qquad H_n^Q := \begin{cases} \gamma(u_{n-1}^h) & \text{for } Q = 1, \\ \gamma(\mathscr{E}_n^2 u^h) & \text{for } Q = 2, \\ \gamma(u_{n-1}^h) & \text{for } Q = 4, \\ \gamma(\bar{u}_n^h) & \text{for } Q = 5. \end{cases}$$

Thus

$$\begin{aligned}(\mathscr{B}_n^Q, \phi_i)_0 &= \frac{1}{2}(G_n^Q \sigma_n^h, \phi_i)_0 + \frac{1}{2}(H_n^Q \sigma_{n-1}^h, \phi_i)_0 \\ &= \frac{1}{2}\sum_{j=1}^{N_V}\Big((G_n^Q \phi_j, \phi_i)_0 \, \boldsymbol{S}_j^n + (H_n^Q \phi_j, \phi_i)_0 \, \boldsymbol{S}_j^{n-1}\Big) \\ &= \frac{1}{2}\sum_{j=1}^{N_V}\Big(\underline{\boldsymbol{X}}_{ij}^{Q,n} \, \boldsymbol{S}_j^n + \underline{\boldsymbol{Y}}_{ij}^{Q,n} \, \boldsymbol{S}_j^{n-1}\Big).\end{aligned} \tag{5.27}$$

For $Q = 3$ we have from the definition of $\mathscr{B}_n^3$ in (5.18) and the definition of the interpolant (2.31),

$$\mathscr{B}_n^3 = \frac{1}{2}\sum_{j=1}^{N_V}\Big(\gamma(\mathscr{E}_n^1 u^h(\boldsymbol{a}_j)) \, \sigma_n^h(\boldsymbol{a}_j) + \gamma(u_{n-1}^h(\boldsymbol{a}_j)) \, \sigma_{n-1}^h(\boldsymbol{a}_j)\Big)\phi_j,$$

thus by using the definition of the diagonal matrices $\underline{\boldsymbol{\Gamma}}^n$ and $\underline{\boldsymbol{\Upsilon}}^n$, we get

$$\mathscr{B}_n^3 = \frac{1}{2}\sum_{j=1}^{N_V}(\underline{\boldsymbol{\Gamma}}_{jj}^n \boldsymbol{S}_j^n + \underline{\boldsymbol{\Upsilon}}_{jj}^n \boldsymbol{S}_j^{n-1})\phi_j.$$

Because of the linearity of $(\cdot, \cdot)_0$, this implies

$$(\mathscr{B}_n^3, \phi_i)_0 = \frac{1}{2}\sum_{j=1}^{N_V}\underline{\boldsymbol{C}}_{ij}(\underline{\boldsymbol{\Gamma}}_{jj}^n \boldsymbol{S}_j^n + \underline{\boldsymbol{\Upsilon}}_{jj}^n \boldsymbol{S}_j^{n-1}).$$

With the definition of $\underline{\boldsymbol{X}}^{3,n}$ and $\underline{\boldsymbol{Y}}^{3,n}$ from (5.24), we obtain for $Q = 3$:

$$(\mathscr{B}_n^Q, \phi_i)_0 = \frac{1}{2}\sum_{j=1}^{N_V}\Big(\underline{\boldsymbol{X}}_{ij}^{Q,n}\ \boldsymbol{S}_j^n + \underline{\boldsymbol{Y}}_{ij}^{Q,n}\ \boldsymbol{S}_j^{n-1}\Big). \tag{5.28}$$

Now recalling (5.27) we see that (5.28) holds in any of the five cases $Q \in \{1, 2, 3, 4, 5\}$. To continue insert (5.28) into (5.26), which gives

$$\begin{aligned}\frac{1}{k}\sum_{j=1}^{N_V}(\phi_j, \phi_i)_0 \cdot (\boldsymbol{S}_j^n - \boldsymbol{S}_j^{n-1}) + \frac{1}{2}\sum_{j=1}^{N_V}\Big(\underline{\boldsymbol{X}}_{ij}^{Q,n}\ \boldsymbol{S}_j^n + \underline{\boldsymbol{Y}}_{ij}^{Q,n}\ \boldsymbol{S}_j^{n-1}\Big)& \\ = \frac{1}{2}\sum_{j=1}^{N_V}(\phi_j, \phi_i)_0 \cdot (\boldsymbol{U}_j^n + \boldsymbol{U}_j^{n-1}) + (\bar{\hat{f}}, \phi_i)_0.&\end{aligned} \tag{5.29}$$

Combining these equations for $i = 1, \ldots, N_V$ into a matrix equation and rearranging gives

$$\begin{aligned}-\tfrac{1}{2}\underline{\boldsymbol{C}} \cdot \boldsymbol{U}^n + (\tfrac{1}{k}\underline{\boldsymbol{C}} + \tfrac{1}{2}\underline{\boldsymbol{X}}^{Q,n}) \cdot \boldsymbol{S}^n& \\ = \tfrac{1}{2}\underline{\boldsymbol{C}} \cdot \boldsymbol{U}^{n-1} + (\tfrac{1}{k}\underline{\boldsymbol{C}} - \tfrac{1}{2}\underline{\boldsymbol{Y}}^{Q,n}) \cdot \boldsymbol{S}^{n-1} + \hat{\boldsymbol{L}}^n.&\end{aligned} \tag{5.30}$$

To conclude the proof, we only have to combine (5.25) and (5.30) to obtain (5.23). □

5.3.1 Matrix form for $Q = 3$

We are now in a position to show why the definition of method $Q = 3$ is useful. From the next lemma, we can see that the matrix formulation can be simplified so that only two linear systems of N_V equations need to be solved instead of the system of $2N_V$ equations of (5.23). In fact, if $\hat{f} = 0$, only (5.31) contains a linear system to be solved.

Lemma 5.8 (Alternative matrix formulation for $Q = 3$).
If $Q = 3$, solving the system (5.23) *is equivalent to solving the following two equations:*

$$\begin{aligned}\Big(\tfrac{1}{k}\underline{\boldsymbol{C}} + \tfrac{1}{2}\underline{\boldsymbol{A}} + \tfrac{1}{4}\underline{\boldsymbol{A}}\underline{\boldsymbol{B}}^n\Big)\boldsymbol{U}^n = \Big(\tfrac{1}{k}\underline{\boldsymbol{C}} - \tfrac{1}{2}\underline{\boldsymbol{A}} - \tfrac{1}{4}\underline{\boldsymbol{A}}\underline{\boldsymbol{B}}^n\Big)\boldsymbol{U}^{n-1}& \\ - \Big(\tfrac{1}{2}\underline{\boldsymbol{A}} + \tfrac{1}{2}\underline{\boldsymbol{A}}\underline{\boldsymbol{B}}^n\underline{\boldsymbol{D}}^n\Big)\boldsymbol{S}^{n-1}& \\ - \tfrac{1}{2}\underline{\boldsymbol{A}}\underline{\boldsymbol{B}}^n\underline{\boldsymbol{C}}^{-1}\hat{\boldsymbol{L}}^n + \boldsymbol{L}^n&\end{aligned} \tag{5.31}$$

$$\boldsymbol{S}^n = \underline{\boldsymbol{B}}^n\Big(\tfrac{1}{2}\boldsymbol{U}^n + \tfrac{1}{2}\boldsymbol{U}^{n-1} + \underline{\boldsymbol{D}}^n\boldsymbol{S}^{n-1} + \underline{\boldsymbol{C}}^{-1}\hat{\boldsymbol{L}}^n\Big), \tag{5.32}$$

where $\underline{\boldsymbol{B}}^n$ and $\underline{\boldsymbol{D}}^n$ are diagonal matrices, defined by

$$\begin{aligned}\underline{\boldsymbol{B}}^n &:= (\tfrac{1}{k}\underline{\boldsymbol{I}} + \tfrac{1}{2}\underline{\boldsymbol{\Gamma}}^n)^{-1},\\ \underline{\boldsymbol{D}}^n &:= (\tfrac{1}{k}\underline{\boldsymbol{I}} - \tfrac{1}{2}\underline{\boldsymbol{\Upsilon}}^n).\end{aligned} \tag{5.33}$$

Here, $\underline{\boldsymbol{I}}$ denotes the identity matrix.

Proof. In the special case of $Q = 3$, (5.30) becomes

$$-\tfrac{1}{2}\underline{\boldsymbol{C}} \cdot \boldsymbol{U}^n + (\tfrac{1}{k}\underline{\boldsymbol{C}} + \tfrac{1}{2}\underline{\boldsymbol{C}}\underline{\boldsymbol{\Gamma}}^n) \cdot \boldsymbol{S}^n = \tfrac{1}{2}\underline{\boldsymbol{C}} \cdot \boldsymbol{U}^{n-1} + (\tfrac{1}{k}\underline{\boldsymbol{C}} - \tfrac{1}{2}\underline{\boldsymbol{C}}\underline{\boldsymbol{\Upsilon}}^n) \cdot \boldsymbol{S}^{n-1} + \hat{\boldsymbol{L}}^n.$$

Pre-multiplying this equation by $\underline{\boldsymbol{C}}^{-1}$ yields

$$-\tfrac{1}{2}\boldsymbol{U}^n + (\tfrac{1}{k}\underline{\boldsymbol{I}} + \tfrac{1}{2}\underline{\boldsymbol{\Gamma}}^n) \cdot \boldsymbol{S}^n = \tfrac{1}{2}\boldsymbol{U}^{n-1} + (\tfrac{1}{k}\underline{\boldsymbol{I}} - \tfrac{1}{2}\underline{\boldsymbol{\Upsilon}}^n) \cdot \boldsymbol{S}^{n-1} + \underline{\boldsymbol{C}}^{-1}\hat{\boldsymbol{L}}^n.$$

By solving this equation for $\boldsymbol{S}^n$ we get (5.32). Now, using (5.32) to replace $\boldsymbol{S}^n$ in (5.25) and rearranging yields (5.31). Finally note that all these steps can be reverted so that (5.31) and (5.32) again imply (5.25) and (5.30) and these two formulations are equivalent. □

5.4 About uniqueness of discrete solutions

In this section we examine the question of whether the discrete scheme of Problem 5.4 actually defines a unique solution. The existence and uniqueness for $Q \in \{1,2\}$ is established in Lemma 5.9. For $Q = 3$, it seems difficult to prove uniqueness, as we will see. For the nonlinear methods $Q \in \{4,5\}$, an example is given that shows that there can be more than one discrete solution.

5.4.1 Methods $Q \in \{1,2\}$

Lemma 5.9 (Unique discrete solution for $Q = 1,2$).
Suppose that $Q \in \{1,2\}$. Then the discrete scheme of Problem 5.4 has a unique solution (u_n^h, σ_n^h) at each time step $n \in \{0,\dots,N\}$.

Proof. The lemma is proved by induction over $n \in \{0,\dots,N\}$.
For $n = 0$, the initial conditions (5.15c) and (5.15d) give the unique solution (u_0^h, σ_0^h). Assume now that up to time step t_{n-1} for some $n > 0$ the discrete solutions were uniquely defined. We need to show that this implies the uniqueness of solution at the next time step t_n. At time step t_n the discrete solution (u_n^h, σ_n^h) is determined by the linear system of equations (5.15a), (5.15b) in a finite-dimensional space. To show uniqueness of the solution (u_n^h, σ_n^h) at this time step, we pick any (hypothetical) two solutions and show that these two solutions must be identical.

Denote the two solutions at time step t_n by (u_n^h, σ_n^h) and $(\tilde{u}_n^h, \tilde{\sigma}_n^h)$. Define their differences $z := u_n^h - \tilde{u}_n^h$ and $p := \sigma_n^h - \tilde{\sigma}_n^h$. Subtracting (5.15a) for the tilde-solution from the same equation for the non-tilde solution and repeating this for (5.15b) gives

$$\tfrac{1}{k}(z,v)_0 + \tfrac{1}{2}(z,v)_V + \tfrac{1}{2}(p,v)_V = 0 \qquad \forall v \in V^h, \tag{5.34a}$$

$$\tfrac{1}{k}(p,v)_0 + (\mathscr{B}_n^Q - \tilde{\mathscr{B}}_n^Q, v)_0 = \tfrac{1}{2}(z,v)_0 \qquad \forall v \in V^h. \tag{5.34b}$$

Here, the tilde version of $\mathscr{B}_n^Q$ is

$$\tilde{\mathscr{B}}_n^Q := \begin{cases} \frac{1}{2}\gamma(\mathscr{E}_n^1 u^h)\tilde{\sigma}_n^h + \frac{1}{2}\gamma(u_{n-1}^h)\sigma_{n-1}^h & \text{for } Q = 1, \\ \frac{1}{2}\gamma(\mathscr{E}_n^2 u^h)(\tilde{\sigma}_n^h + \sigma_{n-1}^h) & \text{for } Q = 2. \end{cases}$$

Choosing $v = \mathcal{G}^h z$ in (5.34a) and $v = p$ in (5.34b) and adding the equations yields

$$\tfrac{1}{k}\|\mathcal{G}^h z\|_V^2 + \tfrac{1}{k}\|p\|_0^2 + \tfrac{1}{2}\|z\|_0^2 + (\mathscr{B}_n^Q - \tilde{\mathscr{B}}_n^Q, p)_0 = 0. \tag{5.35}$$

Observe that for both $Q = 1$ and $Q = 2$

$$\mathscr{B}_n^Q - \tilde{\mathscr{B}}_n^Q = \frac{1}{2}\gamma(\mathscr{E}_n^Q u^h)\, p. \tag{5.36}$$

Thus, in both cases we have that

$$(\mathscr{B}_n^Q - \tilde{\mathscr{B}}_n^Q, p)_0 \geq \frac{\check{\gamma}}{2}\|p\|_0^2,$$

and inserting this in (5.35) gives

$$\tfrac{1}{k}\|\mathcal{G}^h z\|_V^2 + \tfrac{1}{k}\|p\|_0^2 + \tfrac{1}{2}\|z\|_0^2 + \tfrac{\check{\gamma}}{2}\|p\|_0^2 \leq 0,$$

whence $z = 0$ and $p = 0$ and thus the two solutions are identical. In fact, by subtracting the system for two solutions, we have obtained the homogeneous linear system and have shown that this homogeneous system has only the trivial solution $z = p = 0$. Hence the system matrix is invertible and we have also proved the *existence* of a solution. □

5.4.2 Method $Q = 3$

Remark 5.10 (Does $Q = 3$ have a unique solution?).
There are at least two different ways in which one could try to prove uniqueness of solutions for $Q = 3$:

- The first way is to follow the example of the proof of Lemma 5.9. If we can prove that
$$(\mathscr{B}_n^3 - \tilde{\mathscr{B}}_n^3, p)_0 \geq 0, \tag{5.37}$$
then the proof would work in the same way as for $Q = 1$ and 2. Because of (5.36) and the linearity of I^h, we can reformulate the left-hand side of (5.37) as
$$(\mathscr{B}_n^3 - \tilde{\mathscr{B}}_n^3, p)_0 = (I^h[\gamma(\mathscr{E}_n^1 u^h)p], p)_0.$$

- Another possibility is to use the simplified matrix formulation of Lemma 5.8. As the diagonal matrices have only positive entries, they are invertible, so the only problem is to prove that the matrix
$$\tfrac{1}{k}\underline{\boldsymbol{C}} + \tfrac{1}{2}\underline{\boldsymbol{A}} + \tfrac{1}{4}\underline{\boldsymbol{A}}\underline{\boldsymbol{B}}^n$$
from (5.31) is invertible. Now $\underline{\boldsymbol{C}}$ and $\underline{\boldsymbol{A}}$ are positive definite and the sum of positive definite matrices is again positive definite and hence invertible, so it would be sufficient to show that $\underline{\boldsymbol{A}}\underline{\boldsymbol{B}}^n$ is positive definite. This seems to be difficult however. Even though $\underline{\boldsymbol{A}}$ and $\underline{\boldsymbol{B}}^n$ are both symmetric positive definite and $\underline{\boldsymbol{B}}^n$ is also diagonal, this does not imply that $\underline{\boldsymbol{A}}\underline{\boldsymbol{B}}^n$ is positive definite, see Lemma A.30.

In both of these cases the remaining difficulties seem hard to overcome and the question of whether $Q = 3$ must always have a unique solution remains open.

5.4.3 Methods $Q \in \{4, 5\}$

For methods $Q = 4$ and $Q = 5$ the nonlinear system of equations for the discrete solution can have more than one solution. We are going to see this on one particularly simple example, where the nonlinear system can be reduced to a nonlinear equation in just one unknown. This is useful, because simple one-dimensional nonlinear solvers can be used to find *all* solutions. For the implementation of the general case involving many unknowns only iterative schemes are feasible and these will pick just *one* solution, depending on the initial guess.

Example 5.11 (Multiple discrete solutions for $Q \in \{4, 5\}$).
On the one-dimensional domain $\Omega = (0, 1)$ with $\Gamma_D = \{0\}$ and $\Gamma_N = \{1\}$ discretised with only one finite element, i.e. $h = 1$, construct a problem with the exact solution given by
$$u(t, x) = x, \quad \sigma(t, x) = (1 + t)x.$$
For getting this exact solution, we have to choose
$$f(t, x) = 0, \quad g(t, x) = 2 + t, \quad \hat{f}(t, x) = \gamma(x)(1 + t)x.$$
The finite element space is just one-dimensional:
$$V^h = \{x \mapsto ax \ : \ a \in \mathbb{R}\}.$$
Thus we can write
$$u_n^h(x) = U_n\, x, \quad \sigma_n^h(x) = S_n\, x,$$

for some $U_n, S_n \in \mathbb{R}$. For the initial solution we get $U_0 = 1$, $S_0 = 1$. Using these representations for u_n^h and σ_n^h in (5.15a) and in (5.15b) and choosing the basis function $v(x) = x$ gives

$$\tfrac{1}{3k}(U_n - U_{n-1}) + \tfrac{1}{2}(U_n + U_{n-1}) + \tfrac{1}{2}(S_n + S_{n-1}) = 2 + t_{n-1} + \tfrac{k}{2}, \tag{5.38}$$

$$\tfrac{1}{3k}(S_n - S_{n-1}) + \tfrac{1}{2}(S_n \overset{\dashv}{G}{}_n^Q(U_n) + S_{n-1}\overset{\vdash}{G}{}_n^Q(U_n)) = \tfrac{1}{6}(U_n + U_{n-1}) + 1 + t_{n-1} + \tfrac{k}{2}G(1), \tag{5.39}$$

where we have used $(x, x)_0 = 1/3$ and $(x, x)_V = 1$ and where

$$G(y) := \int_0^1 \gamma(yx)x^2\,dx,$$

$$\overset{\dashv}{G}{}_n^Q(y) := \begin{cases} G(y) & \text{if } Q = 4, \\ G(\tfrac{1}{2}(y + U_{n-1})) & \text{if } Q = 5, \end{cases}$$

$$\overset{\vdash}{G}{}_n^Q(y) := \begin{cases} G(U_{n-1}) & \text{if } Q = 4, \\ G(\tfrac{1}{2}(y + U_{n-1})) & \text{if } Q = 5. \end{cases}$$

At each time step this has to be solved for U_n and S_n using the previous discrete solution. In order to get an equation with only one unknown, we can eliminate S_n by solving (5.38) for S_n and substituting this in (5.39). We obtain that U_n is a solution of

$$f(y) = 0,$$

where

$$f(y) = \Big(\tfrac{1}{3k} + \tfrac{1}{2}\overset{\dashv}{G}{}_n^Q(y)\Big) \cdot \Big(-S_{n-1} - (1 - \tfrac{2}{3k})y + 4 + 2t_{n-1} + k\Big) + \Big(-\tfrac{1}{3k} + \tfrac{1}{2}\overset{\dashv}{G}{}_n^Q(y)\Big)S_{n-1} - \tfrac{1}{6}(y + U_{n-1}) - 1 - t_{n-1} - \tfrac{k}{2}G(1).$$

Using numerical integration and a nonlinear solver, we can find all solutions y to this equation.

For both $Q = 4$ and 5 we want to look at the number of solutions we get by varying the time step k. We choose the sine-version of γ (Example 3.8) with $\check{\gamma} = 1$, $\hat{\gamma} = 6$, $\Delta = 0.5$ and $u_a = 1$, go up to end time $T = 1$ and vary the time step $k \in \{1, 0.5\}$.

For method $Q = 4$ and $k = 1$, we obtain the following three discrete solutions, see Figure 5.1:

$$(U_1, S_1) \in \{(0.838743, 2.26876), (1, 2), (1.50775, 1.15375)\}.$$

By decreasing the time step to $k = 0.5$, the solutions become unique for both the first and the second time step, see Figure 5.2:

$$(U_1, S_1) = (1, 1.5) \text{ and } (U_2, S_2) = (1, 2).$$

The same behaviour can be observed for method $Q = 5$. In this case, we also obtain three discrete solutions for $k = 1$, see Figure 5.3:

$$(U_1, S_1) \in \{(0.803998, 2.32667), (1, 2), (1.64412, 0.926461)\}.$$

Again, by decreasing the time step to $k = 0.5$, the solutions become unique for both the first and the second time step, see Figure 5.4:

$$(U_1, S_1) = (1, 1.5) \text{ and } (U_2, S_2) = (1, 2).$$

This example shows that non-uniqueness is possible for $Q \in \{4, 5\}$. We will see this again in Section 7.4.2, where the nonlinear solver seems to converge to the 'wrong' discrete solution.

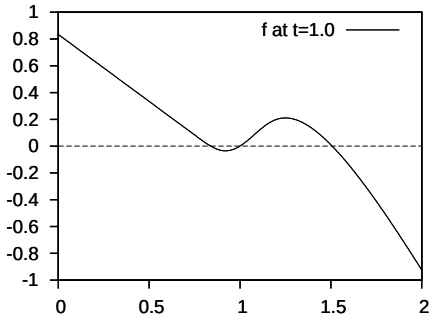

Figure 5.1: Function f from Example 5.11 for $Q = 4$, with $k = 1$ (only one time step). In this case the discrete solution is not unique.

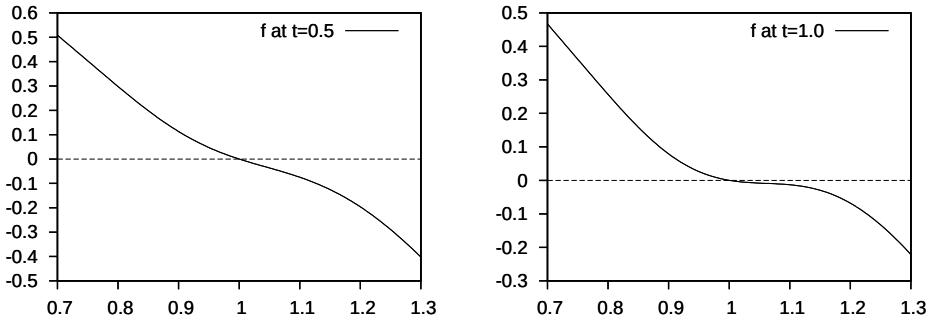

Figure 5.2: Function f from Example 5.11 for $Q = 4$, with $k = 0.5$ (two time steps). Now the discrete solution is unique.

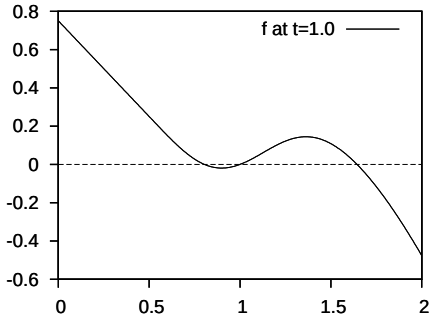

Figure 5.3: Function f for $Q = 5$ from Example 5.11, with $k = 1$ (only one time step). In this case the discrete solution is not unique.

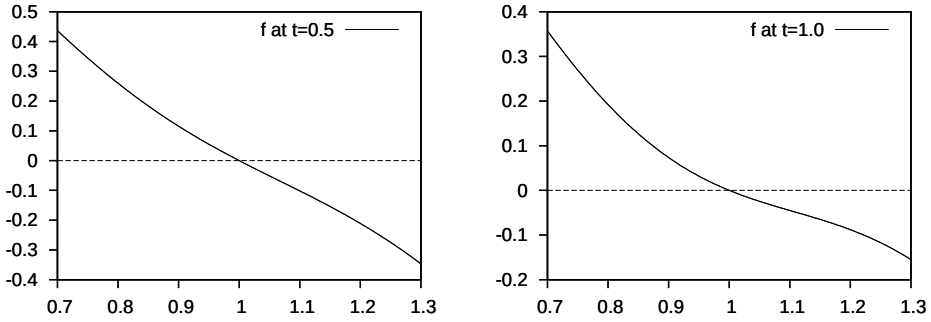

Figure 5.4: Function f from Example 5.11 for $Q = 5$, with $k = 0.5$ (two time steps). Now the discrete solution is unique.

5.5 Boundedness and stability of discrete solution

In this section, we present stability results for the discrete solutions. In Subsection 5.5.1 we give a bound for u_n^h, first using the inverse Laplacian (Proposition 5.14), and then extending this result to stronger norms using the $u+\sigma$-approach (Proposition 5.16).

Then in Subsection 5.5.2 a bound for $\partial_t u_n^h$ is proved.

5.5.1 Bound on solution

Lemma 5.12 (Upper bound for $\mathscr{B}_n^Q$).
For methods $Q = 1, 2, 4, 5$ we have

$$\|\mathscr{B}_n^Q(u^h, \sigma^h)\|_0 \leq \frac{\hat{\gamma}}{2}\left(\|\sigma_n^h\|_0 + \|\sigma_{n-1}^h\|_0\right).$$

Proof. We look at the cases separately. For $Q = 1$:

$$\|\mathscr{B}_n^1\|_0 \leq \frac{1}{2}\left(\|\gamma(\mathscr{E}_n^1 u^h)\sigma_n^h\|_0 + \|\gamma(u_{n-1}^h)\sigma_{n-1}^h\|_0\right) \leq \frac{\hat{\gamma}}{2}\left(\|\sigma_n^h\|_0 + \|\sigma_{n-1}^h\|_0\right).$$

For $Q = 2$:

$$\|\mathscr{B}_n^2\|_0 \leq \hat{\gamma}\|\bar{\sigma}_n^h\|_0 \leq \frac{\hat{\gamma}}{2}\left(\|\sigma_n^h\|_0 + \|\sigma_{n-1}^h\|_0\right).$$

For $Q = 4$:

$$\|\mathscr{B}_n^4\|_0 \leq \frac{1}{2}\left(\|\gamma(u_n^h)\sigma_n^h\|_0 + \|\gamma(u_{n-1}^h)\sigma_{n-1}^h\|_0\right) \leq \frac{\hat{\gamma}}{2}\left(\|\sigma_n^h\|_0 + \|\sigma_{n-1}^h\|_0\right).$$

For $Q = 5$:

$$\|\mathscr{B}_n^5\|_0 \leq \hat{\gamma}\|\bar{\sigma}_n^h\|_0 \leq \frac{\hat{\gamma}}{2}\left(\|\sigma_n^h\|_0 + \|\sigma_{n-1}^h\|_0\right).$$

□

Proving Lemma 5.12 for the remaining method $Q = 3$ appears to be difficult, mainly because the nodal interpolation I^h lacks L_2-stability, see Example 2.30. However, this example does not show that Lemma 5.12 cannot be proved, maybe it can but with a bigger constant.

Lemma 5.13 (Lower bounds for $\mathscr{B}_n^Q$).
For methods $Q \in \{2, 5\}$, we have

$$(\mathscr{B}_n^Q, \bar{\sigma}_n^h)_0 \geq \check{\gamma}\|\bar{\sigma}_n^h\|_0^2. \tag{5.40}$$

For methods $Q \in \{1, 4\}$, and for any $\epsilon > 0$, there is a constant C_ϵ such that

$$(\mathscr{B}_n^Q, \bar{\sigma}_n^h)_0 \geq \check{\gamma}\|\bar{\sigma}_n^h\|_0^2 - \epsilon\|\sigma_n^h\|_0^2 - C_\epsilon\|\sigma_{n-1}^h\|_0^2. \tag{5.41}$$

In fact, $C_\epsilon = \frac{\hat{\gamma}}{2} + \frac{\hat{\gamma}^2}{16\epsilon}$.

Proof. For $Q \in \{2, 5\}$ the assertion follows immediately from the definition of $\mathscr{B}_n^Q$. Thus, for the remainder, let $Q \in \{1, 4\}$. For both these methods, $\mathscr{B}_n^Q$ has the following form:

$$\mathscr{B}_n^Q = \frac{1}{2}\gamma_n^{\dashv}\sigma_n^h + \frac{1}{2}\gamma_n^{\vdash}\sigma_{n-1}^h,$$

with certain functions $\overset{\dashv}{\gamma}_n$ and $\overset{\vdash}{\gamma}_n$, depending on u^h, such that $\check{\gamma} \le \overset{\dashv}{\gamma}_n(\boldsymbol{x}) \le \hat{\gamma}$ and $\check{\gamma} \le \overset{\vdash}{\gamma}_n(\boldsymbol{x}) \le \hat{\gamma}$. For later use, we define $G := \overset{\dashv}{\gamma}_n - \overset{\vdash}{\gamma}_n$ and observe that $\|G\|_{L_\infty(\Omega)} \le 2\hat{\gamma}$. We can rearrange

$$\mathscr{B}_n^Q = \frac{1}{2}\overset{\dashv}{\gamma}_n\sigma_n^h + \frac{1}{2}\overset{\vdash}{\gamma}_n\sigma_{n-1}^h = \frac{1}{2}\overset{\dashv}{\gamma}_n(\sigma_n^h + \sigma_{n-1}^h) - \frac{1}{2}(\overset{\dashv}{\gamma}_n - \overset{\vdash}{\gamma}_n)\sigma_{n-1}^h.$$

Thus

$$\begin{aligned}(\mathscr{B}_n^Q, \bar{\sigma}_n^h)_0 &= (\overset{\dashv}{\gamma}_n\bar{\sigma}_n^h, \bar{\sigma}_n^h)_0 - \frac{1}{2}\big((\overset{\dashv}{\gamma}_n - \overset{\vdash}{\gamma}_n)\sigma_{n-1}^h, \bar{\sigma}_n^h\big)_0 \\ &\ge \check{\gamma}\|\bar{\sigma}_n^h\|_0^2 - \frac{1}{2}|\big((\overset{\dashv}{\gamma}_n - \overset{\vdash}{\gamma}_n)\sigma_{n-1}^h, \bar{\sigma}_n^h\big)_0|.\end{aligned}$$

Using Young's inequality, we get for any $\epsilon_1 > 0$,

$$\begin{aligned}\frac{1}{2}|\big((\overset{\dashv}{\gamma}_n - \overset{\vdash}{\gamma}_n)\sigma_{n-1}^h, \bar{\sigma}_n^h\big)_0| &= \frac{1}{4}|\big(G\sigma_{n-1}^h, \sigma_n^h + \sigma_{n-1}^h\big)_0| \\ &\le \frac{1}{4}|(G\sigma_{n-1}^h, \sigma_n^h)_0| + \frac{1}{4}|(G\sigma_{n-1}^h, \sigma_{n-1}^h)_0| \\ &\le \frac{1}{4}\|G\sigma_{n-1}^h\|_0\,\|\sigma_n^h\|_0 + \frac{\hat{\gamma}}{2}\|\sigma_{n-1}^h\|_0^2 \\ &\le \frac{\hat{\gamma}}{2}(\|\sigma_{n-1}^h\|_0\,\|\sigma_n^h\|_0 + \|\sigma_{n-1}^h\|_0^2) \\ &\le \frac{\hat{\gamma}}{2}\left(\frac{\epsilon_1}{2}\|\sigma_n^h\|_0^2 + (\frac{1}{2\epsilon_1} + 1)\|\sigma_{n-1}^h\|_0^2\right).\end{aligned}$$

We choose $\epsilon_1 = \frac{4}{\hat{\gamma}}\epsilon$ to finish the proof. □

Proposition 5.14 (Boundedness of discrete solutions using inverse Laplacian).
For methods $Q \in \{1, 2, 4, 5\}$ we have, for any $m \in \{1, \dots, N\}$,

$$\begin{aligned}&\|\mathcal{G}^h u_m^h\|_V^2 + \|\sigma_m^h\|_0^2 + k\sum_{n=1}^m \|\bar{u}_n^h\|_0^2 + k\sum_{n=1}^m \|\bar{\sigma}_n^h\|_0^2 \\ &\quad\le C\left(\|\breve{u}\|_{V'}^2 + \|\breve{\sigma}\|_0^2 + \|L\|_{L_\infty(0,T;V')}^2 + \|\hat{f}\|_{L_\infty(0,T;L_2(\Omega))}^2\right).\end{aligned} \tag{5.42}$$

Remark 5.15 (Can the bars be removed?).
As an immediate consequence of Proposition 5.14, we also have the following bound:

$$k\sum_{n=1}^m \|\sigma_n^h\|_0^2 \le C\left(\|\breve{u}\|_{V'}^2 + \|\breve{\sigma}\|_0^2 + \|L\|_{L_\infty(0,T;V')}^2 + \|\hat{f}\|_{L_\infty(0,T;L_2(\Omega))}^2\right) \tag{5.43}$$

but it seems as if

$$k\sum_{n=1}^m \|u_n^h\|_0^2 \le C\left(\|\breve{u}\|_{V'}^2 + \|\breve{\sigma}\|_0^2 + \|L\|_{L_\infty(0,T;V')}^2 + \|\hat{f}\|_{L_\infty(0,T;L_2(\Omega))}^2\right) \tag{5.44}$$

does not follow from this proposition. It does however follow from Proposition 5.16, as we will see later.

Proof of Proposition 5.14. Choosing $v = \mathcal{G}^h\bar{u}_n^h$ in (5.15a) and $v = \bar{\sigma}_n^h$ in (5.15b) and adding the results gives

$$\begin{aligned}&(\partial_t u_n^h, \mathcal{G}^h\bar{u}_n^h)_0 + (\partial_t\sigma_n^h, \bar{\sigma}_n^h)_0 + (\bar{u}_n^h + \bar{\sigma}_n^h, \mathcal{G}^h\bar{u}_n^h)_V + (\mathscr{B}_n^Q, \bar{\sigma}_n^h)_0 \\ &\quad= \langle\bar{L}_n, \mathcal{G}^h\bar{u}_n^h\rangle + (\bar{\hat{f}}_n, \bar{\sigma}_n^h)_0 + (\bar{u}_n^h, \bar{\sigma}_n^h)_0.\end{aligned}$$

Using (5.6), (5.12) and (5.14), this simplifies to

$$\frac{1}{2k}(\|\mathcal{G}^h u_n^h\|_V^2 - \|\mathcal{G}^h u_{n-1}^h\|_V^2) + \frac{1}{2k}(\|\sigma_n^h\|_0^2 - \|\sigma_{n-1}^h\|_0^2) + \|\bar{u}_n^h\|_0^2 + (\mathscr{B}_n^Q, \bar{\sigma}_n^h)_0 \\ = \langle \bar{L}_n, \mathcal{G}^h \bar{u}_n^h \rangle + (\bar{\hat{f}}_n, \bar{\sigma}_n^h)_0.$$

Multiplying by $2k$ and summing for $n = 1, \dots, m$ gives

$$\begin{aligned} &\|\mathcal{G}^h u_m^h\|_V^2 + \|\sigma_m^h\|_0^2 + 2k \sum_{n=1}^m \|\bar{u}_n^h\|_0^2 + 2k \sum_{n=1}^m (\mathscr{B}_n^Q, \bar{\sigma}_n^h)_0 \\ &\qquad = \|\mathcal{G}^h u_0^h\|_V^2 + \|\sigma_0^h\|_0^2 + 2k \sum_{n=1}^m \langle \bar{L}_n, \mathcal{G}^h \bar{u}_n^h \rangle + 2k \sum_{n=1}^m (\bar{\hat{f}}_n, \bar{\sigma}_n^h)_0. \end{aligned} \tag{5.45}$$

Firstly, $\|\sigma_0^h\|_0 \le \|\breve{\sigma}\|_0$. Secondly, to bound $\|\mathcal{G}^h u_0^h\|_V$, we have

$$\|\mathcal{G}^h u_0^h\|_V^2 = (\mathcal{G}^h u_0^h, \mathcal{G}^h u_0^h)_V = (u_0^h, \mathcal{G}^h u_0^h)_0 = (\breve{u}, \mathcal{G}^h u_0^h)_0 \le \|\breve{u}\|_{V'} \, \|\mathcal{G}^h u_0^h\|_V,$$

whence $\|\mathcal{G}^h u_0^h\|_V \le \|\breve{u}\|_{V'}$. Thirdly, we can write

$$\begin{aligned} 2\langle \bar{L}_n, \mathcal{G}^h \bar{u}_n^h \rangle &\le 2\|\bar{L}_n\|_{V'} \, \|\mathcal{G}^h \bar{u}_n^h\|_V \\ &\le \frac{1}{\epsilon} \|\bar{L}_n\|_{V'}^2 + \epsilon \|\mathcal{G}^h \bar{u}_n^h\|_V^2 \\ &\le \frac{1}{\epsilon} \|L\|_{L_\infty(t_{n-1}, t_n; V')}^2 + \epsilon \, C_V^2 \, \|\bar{u}_n^h\|_0^2. \end{aligned}$$

Thus, if we choose $\epsilon := 1/C_V^2$, we get

$$\begin{aligned} 2k \sum_{n=1}^m \langle \bar{L}_n, \mathcal{G}^h \bar{u}_n^h \rangle &\le C_V^2 \, k \sum_{n=1}^m \|L\|_{L_\infty(t_{n-1}, t_n; V')}^2 + k \sum_{n=1}^m \|\bar{u}_n^h\|_0^2 \\ &\le C_V^2 \, T \, \|L\|_{L_\infty(0, T; V')}^2 + k \sum_{n=1}^m \|\bar{u}_n^h\|_0^2. \end{aligned}$$

As a fourth bound we get, using the Cauchy-Schwarz inequality and Young's inequality (A.1),

$$\begin{aligned} 2k \sum_{n=1}^m (\bar{\hat{f}}_n, \bar{\sigma}_n^h)_0 &\le k \sum_{n=1}^m \left(\frac{1}{\check{\gamma}} \|\bar{\hat{f}}_n\|_0^2 + \check{\gamma} \|\bar{\sigma}_n^h\|_0^2 \right) \\ &\le \frac{T}{\check{\gamma}} \|\hat{f}\|_{L_\infty(0, T; L_2(\Omega))}^2 + \check{\gamma} k \sum_{n=1}^m \|\bar{\sigma}_n^h\|_0^2. \end{aligned}$$

Inserting these four bounds in (5.45), the $\|\bar{u}_n^h\|_0^2$-term can be absorbed into the left-hand side and we get

$$\begin{aligned} &\|\mathcal{G}^h u_m^h\|_V^2 + \|\sigma_m^h\|_0^2 + k \sum_{n=1}^m \|\bar{u}_n^h\|_0^2 + 2k \sum_{n=1}^m (\mathscr{B}_n^Q, \bar{\sigma}_n^h)_0 \\ &\qquad \le \|\breve{u}\|_{V'}^2 + \|\breve{\sigma}\|_0^2 + C_V^2 \, T \, \|L\|_{L_\infty(0, T; V')}^2 \\ &\qquad\quad + \frac{T}{\check{\gamma}} \|\hat{f}\|_{L_\infty(0, T; L_2(\Omega))}^2 + \check{\gamma} k \sum_{n=1}^m \|\bar{\sigma}_n^h\|_0^2. \end{aligned} \tag{5.46}$$

For methods $Q = 2$ or 5, we use (5.40), subtract $\check{\gamma}k\sum_{n=1}^{m}\|\bar{\sigma}_n^h\|_0^2$ from both sides and obtain

$$\begin{aligned}\|\mathcal{G}^h u_m^h\|_V^2 + \|\sigma_m^h\|_0^2 + k\sum_{n=1}^{m}\|\bar{u}_n^h\|_0^2 + \check{\gamma}k\sum_{n=1}^{m}\|\bar{\sigma}_n^h\|_0^2 \\ \leq \|\check{u}\|_{V'}^2 + \|\check{\sigma}\|_0^2 + C_V^2\,T\,\|L\|_{L_\infty(0,T;V')}^2 + \frac{T}{\check{\gamma}}\|\hat{f}\|_{L_\infty(0,T;L_2(\Omega))}^2,\end{aligned} \tag{5.47}$$

which concludes the proof for $Q = 2$ or 5.

Now, for methods $Q = 1$ or 4, we use (5.41) in (5.46), again subtract $\check{\gamma}k\sum_{n=1}^{m}\|\bar{\sigma}_n^h\|_0^2$ from both sides and obtain

$$\begin{aligned}\|\mathcal{G}^h u_m^h\|_V^2 + \|\sigma_m^h\|_0^2 + k\sum_{n=1}^{m}\|\bar{u}_n^h\|_0^2 + \check{\gamma}k\sum_{n=1}^{m}\|\bar{\sigma}_n^h\|_0^2 \\ \leq \|\check{u}\|_{V'}^2 + \|\check{\sigma}\|_0^2 + C_V^2\,T\,\|L\|_{L_\infty(0,T;V')}^2 + \frac{T}{\check{\gamma}}\|\hat{f}\|_{L_\infty(0,T;L_2(\Omega))}^2 \\ + 2k\sum_{n=1}^{m}(\epsilon\|\sigma_n^h\|_0^2 + C_\epsilon\|\sigma_{n-1}^h\|_0^2).\end{aligned} \tag{5.48}$$

We write

$$\begin{aligned}2k\sum_{n=1}^{m}(\epsilon\|\sigma_n^h\|_0^2 + C_\epsilon\|\sigma_{n-1}^h\|_0^2) &\leq 2k\epsilon\|\sigma_m^h\|_0^2 + 2(\epsilon + C_\epsilon)\,k\sum_{n=1}^{m-1}\|\sigma_n^h\|_0^2 + 2C_\epsilon\,k\|\sigma_0^h\|_0^2 \\ &\leq 2T\epsilon\|\sigma_m^h\|_0^2 + 2(\epsilon + C_\epsilon)\,k\sum_{n=1}^{m-1}\|\sigma_n^h\|_0^2 + 2TC_\epsilon\|\check{\sigma}\|_0^2.\end{aligned}$$

Choosing $\epsilon = 1/(4T)$ and inserting this into (5.48) gives

$$\begin{aligned}\|\mathcal{G}^h u_m^h\|_V^2 + \frac{1}{2}\|\sigma_m^h\|_0^2 + k\sum_{n=1}^{m}\|\bar{u}_n^h\|_0^2 + \check{\gamma}k\sum_{n=1}^{m}\|\bar{\sigma}_n^h\|_0^2 \\ \leq \|\check{u}\|_{V'}^2 + c_1\|\check{\sigma}\|_0^2 + C_V^2\,T\,\|L\|_{L_\infty(0,T;V')}^2 + \frac{T}{\check{\gamma}}\|\hat{f}\|_{L_\infty(0,T;L_2(\Omega))}^2 + c_2\,k\sum_{n=1}^{m-1}\|\sigma_n^h\|_0^2,\end{aligned}$$

where $c_1 = 1 + 2TC_\epsilon = 1 + \hat{\gamma}T + \hat{\gamma}^2T^2/2$ and $c_2 = 2\epsilon + 2C_\epsilon = 1/(2T) + \hat{\gamma} + \hat{\gamma}^2T/2$. Finally, we use the discrete Grönwall lemma (Lemma A.4), to eliminate the last term on the right-hand side. We get

$$\begin{aligned}\|\mathcal{G}^h u_m^h\|_V^2 + \frac{1}{2}\|\sigma_m^h\|_0^2 + k\sum_{n=1}^{m}\|\bar{u}_n^h\|_0^2 + \check{\gamma}k\sum_{n=1}^{m}\|\bar{\sigma}_n^h\|_0^2 \\ \leq \exp(2c_2T)\left(\|\check{u}\|_{V'}^2 + c_1\|\check{\sigma}\|_0^2 + C_V^2\,T\,\|L\|_{L_\infty(0,T;V')}^2 + \frac{T}{\check{\gamma}}\|\hat{f}\|_{L_\infty(0,T;L_2(\Omega))}^2\right),\end{aligned}$$

which concludes the proof for $Q \in \{1, 4\}$. □

Proposition 5.16 (Boundedness of discrete solutions using $u + \sigma$-terms). *For methods $Q = 1, 2, 4, 5$ we have for $m \in \{1, \ldots, N\}$,*

$$\begin{aligned}\|u_m^h\|_0^2 + \|\sigma_m^h\|_0^2 + k\sum_{n=1}^{m}\|\bar{u}_n^h + \bar{\sigma}_n^h\|_V^2 \\ \leq C\left(\|\check{u}\|_0^2 + \|\check{\sigma}\|_0^2 + \|L\|_{L_\infty(0,T;V')}^2 + \|\hat{f}\|_{L_\infty(0,T;L_2(\Omega))}^2\right).\end{aligned} \tag{5.49}$$

Proof. Adding (5.15a) and (5.15b) and choosing $v = \bar{u}_n^h + \bar{\sigma}_n^h$ gives

$$\begin{aligned}&\frac{1}{2k}\|u_n^h + \sigma_n^h\|_0^2 - \frac{1}{2k}\|u_{n-1}^h + \sigma_{n-1}^h\|_0^2 + \|\bar{u}_n^h + \bar{\sigma}_n^h\|_V^2 \\ &= -(\mathscr{B}_n^Q, \bar{u}_n^h + \bar{\sigma}_n^h)_0 + \langle \bar{L}_n, \bar{u}_n^h + \bar{\sigma}_n^h\rangle + (\bar{\hat{f}}_n, \bar{u}_n^h + \bar{\sigma}_n^h)_0 + (\bar{u}_n^h, \bar{u}_n^h + \bar{\sigma}_n^h)_0.\end{aligned}$$

Applying $\langle w, v\rangle \leq \|w\|_{V'} \cdot \|v\|_V$ and Young's inequality yields

$$\begin{aligned}&\frac{1}{2k}\|u_n^h + \sigma_n^h\|_0^2 - \frac{1}{2k}\|u_{n-1}^h + \sigma_{n-1}^h\|_0^2 + \|\bar{u}_n^h + \bar{\sigma}_n^h\|_V^2 \\ &\leq 2(\|\mathscr{B}_n^Q\|_{V'}^2 + \|\bar{L}_n\|_{V'}^2 + \|\bar{\hat{f}}_n\|_{V'}^2 + \|\bar{u}_n^h\|_{V'}^2) + 4 \cdot \frac{1}{8}\|\bar{u}_n^h + \bar{\sigma}_n^h\|_V^2.\end{aligned}$$

Subtracting $\frac{1}{2}\|\bar{u}_n^h + \bar{\sigma}_n^h\|_V^2$ and multiplying by $2k$ gives

$$\begin{aligned}&\|u_n^h + \sigma_n^h\|_0^2 - \|u_{n-1}^h + \sigma_{n-1}^h\|_0^2 + k\|\bar{u}_n^h + \bar{\sigma}_n^h\|_V^2 \\ &\qquad \leq 4k\left(\|\mathscr{B}_n^Q\|_{V'}^2 + \|\bar{L}_n\|_{V'}^2 + \|\bar{\hat{f}}_n\|_{V'}^2 + \|\bar{u}_n^h\|_{V'}^2\right).\end{aligned} \tag{5.50}$$

Now choosing $v = \bar{\sigma}_n^h$ in (5.15b) gives

$$\frac{1}{2k}\|\sigma_n^h\|_0^2 - \frac{1}{2k}\|\sigma_{n-1}^h\|_0^2 = -(\mathscr{B}_n^Q, \bar{\sigma}_n^h)_0 + (\bar{u}_n^h, \bar{\sigma}_n^h)_0 + (\bar{\hat{f}}_n, \bar{\sigma}_n^h)_0.$$

Multiplying by $2k$, applying the Cauchy-Schwarz inequality and Young's inequality yields

$$\|\sigma_n^h\|_0^2 - \|\sigma_{n-1}^h\|_0^2 \leq k\|\mathscr{B}_n^Q\|_0^2 + k\|\bar{\hat{f}}_n\|_0^2 + k\|\bar{u}_n^h\|_0^2 + 3k\|\bar{\sigma}_n^h\|_0^2. \tag{5.51}$$

Now add (5.50) and (5.51), use $\|v\|_{V'} \leq C_V\|v\|_0$ and obtain

$$\begin{aligned}&\|u_n^h + \sigma_n^h\|_0^2 - \|u_{n-1}^h + \sigma_{n-1}^h\|_0^2 + \|\sigma_n^h\|_0^2 - \|\sigma_{n-1}^h\|_0^2 + k\|\bar{u}_n^h + \bar{\sigma}_n^h\|_V^2 \\ &\qquad \leq 4k\|\bar{L}_n\|_{V'}^2 + (4C_V^2 + 1)k\|\bar{\hat{f}}_n\|_0^2 + (4C_V^2 + 1)k\|\mathscr{B}_n^Q\|_0^2 \\ &\qquad\quad + (4C_V^2 + 1)k\|\bar{u}_n^h\|_0^2 + 3k\|\bar{\sigma}_n^h\|_0^2.\end{aligned}$$

Summing for $n = 1, \ldots, m$ gives

$$\begin{aligned}&\|u_m^h + \sigma_m^h\|_0^2 + \|\sigma_m^h\|_0^2 + k\sum_{n=1}^m \|\bar{u}_n^h + \bar{\sigma}_n^h\|_V^2 \\ &\qquad \leq \|u_0^h + \sigma_0^h\|_0^2 + \|\sigma_0^h\|_0^2 + 4k\sum_{n=1}^m \|\bar{L}_n\|_{V'}^2 + (4C_V^2 + 1)k\sum_{n=1}^m \|\bar{\hat{f}}_n\|_0^2 \\ &\qquad\quad + (4C_V^2 + 1)k\sum_{n=1}^m \|\mathscr{B}_n^Q\|_0^2 + (4C_V^2 + 1)k\sum_{n=1}^m \|\bar{u}_n^h\|_0^2 + 3k\sum_{n=1}^m \|\bar{\sigma}_n^h\|_0^2.\end{aligned}$$

By using (4.7) we obtain

$$\begin{aligned}&\frac{1}{3}\|u_m^h\|_0^2 + \frac{1}{3}\|\sigma_m^h\|_0^2 + k\sum_{n=1}^m \|\bar{u}_n^h + \bar{\sigma}_n^h\|_V^2 \\ &\qquad \leq 3\|u_0^h\|_0^2 + 3\|\sigma_0^h\|_0^2 + 4k\sum_{n=1}^m \|\bar{L}_n\|_{V'}^2 + (4C_V^2 + 1)k\sum_{n=1}^m \|\bar{\hat{f}}_n\|_0^2 \\ &\qquad\quad + (4C_V^2 + 1)k\sum_{n=1}^m \|\mathscr{B}_n^Q\|_0^2 + (4C_V^2 + 1)k\sum_{n=1}^m \|\bar{u}_n^h\|_0^2 + 3k\sum_{n=1}^m \|\bar{\sigma}_n^h\|_0^2.\end{aligned}$$

By using the bound on $\mathscr{B}_n^Q$ from Lemma 5.12 we obtain

$$\begin{aligned}\|u_m^h\|_0^2 + \|\sigma_m^h\|_0^2 + k\sum_{n=1}^m \|\bar{u}_n^h + \bar{\sigma}_n^h\|_V^2 \\ \leq C\Bigg(\|u_0^h\|_0^2 + \|\sigma_0^h\|_0^2 + k\sum_{n=1}^m \|\bar{L}_n\|_{V'}^2 + k\sum_{n=1}^m \|\bar{\hat{f}}_n\|_0^2 \\ + k\sum_{n=1}^m \|\bar{u}_n^h\|_0^2 + k\sum_{n=1}^m \|\bar{\sigma}_n^h\|_0^2 + k\sum_{n=1}^m \|\sigma_n^h\|_0^2 \Bigg).\end{aligned} \tag{5.52}$$

To bound the term involving L, observe that

$$k\sum_{n=1}^m \|\bar{L}_n\|_{V'}^2 \leq k\sum_{n=1}^m \|L\|_{L_\infty(0,T;V')}^2 \leq T\|L\|_{L_\infty(0,T;V')}^2.$$

Similarly, for the term involving $\hat{f}$, we have that

$$k\sum_{n=1}^m \|\bar{\hat{f}}_n\|_0^2 \leq T\|\hat{f}\|_{L_\infty(0,T;L_2(\Omega))}^2.$$

Also note that from (5.15c) and (5.15d), $\|u_0^h\|_0 \leq \|\breve{u}\|_0$ and $\|\sigma_0^h\|_0 \leq \|\breve{\sigma}\|_0$. Using all these in (5.52) gives

$$\begin{aligned}\|u_m^h\|_0^2 + \|\sigma_m^h\|_0^2 + k\sum_{n=1}^m \|\bar{u}_n^h + \bar{\sigma}_n^h\|_V^2 \\ \leq C\Bigg(\|\breve{u}\|_0^2 + \|\breve{\sigma}\|_0^2 + \|L\|_{L_\infty(0,T;V')}^2 + \|\hat{f}\|_{L_\infty(0,T;L_2(\Omega))}^2 \\ + k\sum_{n=1}^m \|\bar{u}_n^h\|_0^2 + k\sum_{n=1}^m \|\bar{\sigma}_n^h\|_0^2 + k\sum_{n=1}^m \|\sigma_n^h\|_0^2 \Bigg).\end{aligned} \tag{5.53}$$

From Proposition 5.14 and Remark 5.15 we get bounds for the terms

$$k\sum_{n=1}^m \|\bar{u}_n^h\|_0^2 + k\sum_{n=1}^m \|\bar{\sigma}_n^h\|_0^2 + k\sum_{n=1}^m \|\sigma_n^h\|_0^2.$$

Inserting these bounds in (5.53) we get

$$\begin{aligned}\|u_m^h\|_0^2 + \|\sigma_m^h\|_0^2 + k\sum_{n=1}^m \|\bar{u}_n^h + \bar{\sigma}_n^h\|_V^2 \\ \leq C\left(\|\breve{u}\|_0^2 + \|\breve{u}\|_{V'}^2 + \|\breve{\sigma}\|_0^2 + \|L\|_{L_\infty(0,T;V')}^2 + \|\hat{f}\|_{L_\infty(0,T;L_2(\Omega))}^2 \right).\end{aligned}$$

Finally note that $\|\breve{u}\|_{V'} \leq C_V\|\breve{u}\|_0$, which concludes the proof. □

As a corollary we note that the sum of $\|u_n^h\|_0^2$ is also bounded:

Corollary 5.17 (Boundedness of discrete solutions).
For methods $Q = 1, 2, 4, 5$, and for any $m \in \{1, \ldots, N\}$,

$$\begin{aligned}\|u_m^h\|_0^2 + \|\sigma_m^h\|_0^2 + k\sum_{n=1}^m \|u_n^h\|_0^2 + k\sum_{n=1}^m \|\sigma_n^h\|_0^2 + k\sum_{n=1}^m \|\bar{u}_n^h + \bar{\sigma}_n^h\|_V^2 \\ \leq C\left(\|\breve{u}\|_0^2 + \|\breve{\sigma}\|_0^2 + \|L\|_{L_\infty(0,T;V')}^2 + \|\hat{f}\|_{L_\infty(0,T;L_2(\Omega))}^2 \right).\end{aligned} \tag{5.54}$$

Proof. From Proposition 5.16, we get for any $n \in \{1, \dots, N\}$

$$\|u_n^h\|_0^2 + \|\sigma_n^h\|_0^2 \leq C \left(\|\breve{u}\|_0^2 + \|\breve{\sigma}\|_0^2 + \|L\|_{L_\infty(0,T;V')}^2 + \|\hat{f}\|_{L_\infty(0,T;L_2(\Omega))}^2 \right).$$

Since $k \sum_{n=1}^m 1 = km \leq T = \text{const}$, this implies

$$k \sum_{n=1}^m \|u_n^h\|_0^2 + k \sum_{n=1}^m \|\sigma_n^h\|_0^2 \leq C \left(\|\breve{u}\|_0^2 + \|\breve{\sigma}\|_0^2 + \|L\|_{L_\infty(0,T;V')}^2 + \|\hat{f}\|_{L_\infty(0,T;L_2(\Omega))}^2 \right),$$

which, combined again with Proposition 5.16, is what we want to prove. □

5.5.2 Bound on differences

Proposition 5.18 (Bound on differences).
For methods $Q = 1, 2, 4, 5$, we have for any $m \in \{1, \dots, N\}$:

$$\begin{aligned} k \sum_{n=1}^m \|\partial_t u_n^h\|_0^2 + k \sum_{n=1}^m \|\partial_t \sigma_n^h\|_0^2 + \|u_m^h + \sigma_m^h\|_V^2 \\ \leq C \left(\|\breve{u}\|_V^2 + \|\breve{\sigma}\|_V^2 + \|L\|_{L_\infty(0,T;L_2(\Omega))}^2 + \|\hat{f}\|_{L_\infty(0,T;L_2(\Omega))}^2 \right). \end{aligned}$$

Proof. Adding (5.15a) and (5.15b) and choosing $v = \partial_t u_n^h + \partial_t \sigma_n^h$ gives

$$\begin{aligned} \|\partial_t u_n^h + \partial_t \sigma_n^h\|_0^2 + (\bar{u}_n^h + \bar{\sigma}_n^h, \partial_t u_n^h + \partial_t \sigma_n^h)_V + (\mathscr{B}_n^Q, \partial_t u_n^h + \partial_t \sigma_n^h)_0 \\ = (\bar{L}_n, \partial_t u_n^h + \partial_t \sigma_n^h)_0 + (\bar{\hat{f}}_n, \partial_t u_n^h + \partial_t \sigma_n^h)_0 + (\bar{u}_n^h, \partial_t u_n^h + \partial_t \sigma_n^h)_0. \end{aligned}$$

Rearranging and observing that

$$(\bar{u}_n^h + \bar{\sigma}_n^h, \partial_t u_n^h + \partial_t \sigma_n^h)_V = \frac{1}{2k} \|u_n^h + \sigma_n^h\|_V^2 - \frac{1}{2k} \|u_{n-1}^h + \sigma_{n-1}^h\|_V^2,$$

we obtain

$$\begin{aligned} \|\partial_t u_n^h + \partial_t \sigma_n^h\|_0^2 + \frac{1}{2k} \|u_n^h + \sigma_n^h\|_V^2 - \frac{1}{2k} \|u_{n-1}^h + \sigma_{n-1}^h\|_V^2 \\ = (\bar{L}_n, \partial_t u_n^h + \partial_t \sigma_n^h)_0 + (\bar{\hat{f}}_n, \partial_t u_n^h + \partial_t \sigma_n^h)_0 \\ - (\mathscr{B}_n^Q, \partial_t u_n^h + \partial_t \sigma_n^h)_0 + (\bar{u}_n^h, \partial_t u_n^h + \partial_t \sigma_n^h)_0. \end{aligned}$$

Now use the Cauchy-Schwarz inequality and Young's inequality, which yields

$$\begin{aligned} \|\partial_t u_n^h + \partial_t \sigma_n^h\|_0^2 + \frac{1}{2k} \|u_n^h + \sigma_n^h\|_V^2 - \frac{1}{2k} \|u_{n-1}^h + \sigma_{n-1}^h\|_V^2 \\ \leq \frac{1}{2\epsilon_1} \|\bar{L}_n\|_0^2 + \frac{1}{2\epsilon_2} \|\bar{\hat{f}}_n\|_0^2 + \frac{1}{2\epsilon_3} \|\mathscr{B}_n^Q\|_0^2 + \frac{1}{2\epsilon_4} \|\bar{u}_n^h\|_0^2 \\ + \frac{1}{2}(\epsilon_1 + \epsilon_2 + \epsilon_3 + \epsilon_4) \|\partial_t u_n^h + \partial_t \sigma_n^h\|_0^2. \end{aligned}$$

Choosing $\epsilon_1 = \epsilon_2 = \epsilon_3 = \epsilon_4 = 1/4$ gives

$$\begin{aligned} \frac{1}{2} \|\partial_t u_n^h + \partial_t \sigma_n^h\|_0^2 + \frac{1}{2k} \|u_n^h + \sigma_n^h\|_V^2 - \frac{1}{2k} \|u_{n-1}^h + \sigma_{n-1}^h\|_V^2 \\ \leq 2\|\bar{L}_n\|_0^2 + 2\|\bar{\hat{f}}_n\|_0^2 + 2\|\mathscr{B}_n^Q\|_0^2 + 2\|\bar{u}_n^h\|_0^2. \end{aligned} \tag{5.55}$$

Now choosing $v = \partial_t \sigma_n^h$ in (5.15b) gives

$$\|\partial_t \sigma_n^h\|_0^2 + (\mathscr{B}_n^Q, \partial_t \sigma_n^h)_0 = (\bar{u}_n^h, \partial_t \sigma_n^h)_0 + (\bar{\hat{f}}_n, \partial_t \sigma_n^h)_0.$$

Rearranging and using Cauchy-Schwarz and Young's inequalities, we obtain

$$\|\partial_t \sigma_n^h\|_0^2 \leq \frac{1}{2\epsilon_1}\|\bar{\hat{f}}_n\|_0^2 + \frac{1}{2\epsilon_2}\|\mathscr{B}_n^Q\|_0^2 + \frac{1}{2\epsilon_3}\|\bar{u}_n^h\|_0^2 + \frac{1}{2}(\epsilon_1+\epsilon_2+\epsilon_3)\|\partial_t \sigma_n^h\|_0^2.$$

Choosing $\epsilon_1 = \epsilon_2 = \epsilon_3 = 1/3$ gives

$$\frac{1}{2}\|\partial_t \sigma_n^h\|_0^2 \leq \frac{3}{2}\big(\|\bar{\hat{f}}_n\|_0^2 + \|\mathscr{B}_n^Q\|_0^2 + \|\bar{u}_n^h\|_0^2\big). \tag{5.56}$$

Multiplying (5.55) by $2k$, multiplying (5.56) by $2k$ and adding the results gives

$$\begin{aligned} &k\|\partial_t u_n^h + \partial_t \sigma_n^h\|_0^2 + k\|\partial_t \sigma_n^h\|_0^2 + \|u_n^h + \sigma_n^h\|_V^2 - \|u_{n-1}^h + \sigma_{n-1}^h\|_V^2 \\ &\quad \leq 4k\|\bar{L}_n\|_0^2 + 7k\|\bar{\hat{f}}_n\|_0^2 + 7k\|\mathscr{B}_n^Q\|_0^2 + 7k\|\bar{u}_n^h\|_0^2. \end{aligned}$$

Using (4.7) this becomes

$$\begin{aligned} &\frac{1}{3}k\|\partial_t u_n^h\|_0^2 + \frac{1}{3}k\|\partial_t \sigma_n^h\|_0^2 + \|u_n^h + \sigma_n^h\|_V^2 - \|u_{n-1}^h + \sigma_{n-1}^h\|_V^2 \\ &\quad \leq 4k\|\bar{L}_n\|_0^2 + 7k\|\bar{\hat{f}}_n\|_0^2 + 7k\|\mathscr{B}_n^Q\|_0^2 + 7k\|\bar{u}_n^h\|_0^2. \end{aligned}$$

Summing for $n = 1, \ldots, m$ yields

$$\begin{aligned} &\frac{1}{3}k\sum_{n=1}^m \|\partial_t u_n^h\|_0^2 + \frac{1}{3}k\sum_{n=1}^m \|\partial_t \sigma_n^h\|_0^2 + \|u_m^h + \sigma_m^h\|_V^2 \\ &\quad \leq \|u_0^h + \sigma_0^h\|_V^2 + 4k\sum_{n=1}^m \|\bar{L}_n\|_0^2 + 7k\sum_{n=1}^m \|\bar{\hat{f}}_n\|_0^2 + 7k\sum_{n=1}^m \|\mathscr{B}_n^Q\|_0^2 + 7k\sum_{n=1}^m \|\bar{u}_n^h\|_0^2. \end{aligned}$$

Using (4.6) to split the term $\|u_0^h + \sigma_0^h\|_V^2$ and applying Lemma 5.12 gives

$$\begin{aligned} &k\sum_{n=1}^m \|\partial_t u_n^h\|_0^2 + k\sum_{n=1}^m \|\partial_t \sigma_n^h\|_0^2 + \|u_m^h + \sigma_m^h\|_V^2 \\ &\quad \leq C\bigg(\|u_0^h\|_V^2 + \|\sigma_0^h\|_V^2 + k\sum_{n=1}^m \|\bar{L}_n\|_0^2 + k\sum_{n=1}^m \|\bar{\hat{f}}_n\|_0^2 \\ &\qquad + k\sum_{n=1}^m \|\bar{u}_n^h\|_0^2 + k\sum_{n=0}^m \|\sigma_n^h\|_0^2\bigg). \end{aligned}$$

Now, we use Corollary 5.17 and we bound the terms involving L and $\hat{f}$. This gives

$$\begin{aligned} &k\sum_{n=1}^m \|\partial_t u_n^h\|_0^2 + k\sum_{n=1}^m \|\partial_t \sigma_n^h\|_0^2 + \|u_m^h + \sigma_m^h\|_V^2 \\ &\quad \leq C\Big(\|u_0^h\|_V^2 + \|\sigma_0^h\|_V^2 + \|\breve{u}\|_0^2 + \|\breve{\sigma}\|_0^2 \\ &\qquad + \|L\|_{L_\infty(0,T;L_2(\Omega))}^2 + \|L\|_{L_\infty(0,T;V')}^2 + \|\hat{f}\|_{L_\infty(0,T;L_2(\Omega))}^2\Big). \end{aligned}$$

To finish the proof, observe that $\|L\|_{L_\infty(0,T;V')} \leq C_V\|L\|_{L_\infty(0,T;L_2(\Omega))}$ and that by Lemma 2.33, $\|u_0^h\|_V \leq C\|\breve{u}\|_V$ and $\|\sigma_0^h\|_V \leq C\|\breve{\sigma}\|_V$. □

Chapter 6

Error Estimate

This chapter presents an a priori error estimate for the Cohen model. The result is proved only for four of the five methods that were presented in the previous chapter, namely for methods $Q \in \{1, 2, 4, 5\}$. The way the error estimate is proved is based on the error analysis for the modified model by Rivière and Shaw in [RS06], see also page 35 in Section 3.3. For that model the estimates could be derived without using the inverse Laplacian or the combined variable $u + \sigma$. It was possible to combine the two equations directly to prove the error estimate. Here however, the approach of calculating with $u + \sigma$ and the approach of using the inverse Laplacian are combined to achieve similar estimates to those in [RS06].

In Section 6.1 some terms involving an elliptic projection are defined and bounds for them are given. These bounds will be needed throughout this chapter.

In Section 6.2 two inequalities are proved that deal just with the linear parts of the system. Therefore, these are general results that still apply to all five methods.

Section 6.3 gives bounds for the nonlinearity error, i.e. for the difference between the nonlinearity $\gamma(u)\sigma$ and its approximation $\mathscr{B}_n^Q$. Here we restrict ourselves to the four methods $Q \in \{1, 2, 4, 5\}$.

In Section 6.4 the error estimate is stated and proved by combining the error inequalities from Section 6.2 with the nonlinearity errors from Section 6.3.

Finally, Section 6.5 contains notes on why it seems more difficult to prove an error estimate for method $Q = 3$.

6.1 Preparations for the error estimate

For proving the error estimate, we need the elliptic projection $u_n^*, \sigma_n^* \in V^h$ of the exact solution, which is defined by

$$u_n^* := \Pi_V^h u_n, \qquad \sigma_n^* := \Pi_V^h \sigma_n, \tag{6.1}$$

and define the differences,

$$\begin{aligned} \chi_n &:= u_n^h - u_n^*, & \eta_n &:= \sigma_n^h - \sigma_n^*, \\ \xi_n &:= u_n - u_n^*, & \theta_n &:= \sigma_n - \sigma_n^*. \end{aligned} \tag{6.2}$$

We now need a version of Lemma 2.29 for time dependent functions:

Lemma 6.1 (Approximation error for elliptic projection).
Suppose that $\mathcal{ER}^q$ holds and that $v \in H^m(0, T; V \cap H^{s+1}(\Omega))$ for some $m \in \mathbb{N}$ and $s \in \{1, \dots, r\}$. Then the elliptic projection $v^* := \Pi_V^h v$ is in $H^m(0, T; V^h \cap H^{s+1}(\Omega))$

and

$$\left\| \frac{\partial^m}{\partial t^m}(v(t) - v^*(t)) \right\|_V \leq Ch^s \left\| \frac{\partial^m}{\partial t^m} v(t) \right\|_{H^{s+1}(\Omega)} \qquad \forall t \in (0,T) \text{ a.e.},$$

and

$$\left\| \frac{\partial^m}{\partial t^m}(v(t) - v^*(t)) \right\|_0 \leq Ch^{s+q} \left\| \frac{\partial^m}{\partial t^m} v(t) \right\|_{H^{s+1}(\Omega)} \qquad \forall t \in (0,T) \text{ a.e.}$$

Proof. This is Lemma 2.34, applied to time-dependent functions. Observe that the operators for elliptic projection Π_V^h and differentiation with respect to time commute. □

Lemma 6.2 (Bounds involving ξ_n and θ_n).
Suppose that $\mathcal{ER}^q$ holds. Whenever the exact solutions u and σ have the indicated regularity, we have

$$\|\xi_n\|_V \leq Ch^r \|u_n\|_{H^{r+1}(\Omega)}, \tag{6.3}$$

$$\|\theta_n\|_V \leq Ch^r \|\sigma_n\|_{H^{r+1}(\Omega)}, \tag{6.4}$$

$$\|\xi_n\|_0 \leq Ch^{r+q} \|u_n\|_{H^{r+1}(\Omega)}, \tag{6.5}$$

$$\|\theta_n\|_0 \leq Ch^{r+q} \|\sigma_n\|_{H^{r+1}(\Omega)}, \tag{6.6}$$

$$k \sum_{n=1}^m \|\bar{\xi}_n\|_V^2 \leq Ch^{2r} \|u\|^2_{L_\infty(0,T;H^{r+1}(\Omega))}, \tag{6.7}$$

$$k \sum_{n=1}^m \|\bar{\theta}_n\|_V^2 \leq Ch^{2r} \|\sigma\|^2_{L_\infty(0,T;H^{r+1}(\Omega))}, \tag{6.8}$$

$$k \sum_{n=1}^m \|\bar{\xi}_n\|_0^2 \leq Ch^{2(r+q)} \|u\|^2_{L_\infty(0,T;H^{r+1}(\Omega))}, \tag{6.9}$$

$$k \sum_{n=1}^m \|\bar{\theta}_n\|_0^2 \leq Ch^{2(r+q)} \|\sigma\|^2_{L_\infty(0,T;H^{r+1}(\Omega))}, \tag{6.10}$$

$$\|\chi_0\|_0 \leq \|\xi_0\|_0 \leq Ch^{r+q} \|\breve{u}\|_{H^{r+1}(\Omega)}, \tag{6.11}$$

$$\|\eta_0\|_0 \leq \|\theta_0\|_0 \leq Ch^{r+q} \|\breve{\sigma}\|_{H^{r+1}(\Omega)}, \tag{6.12}$$

$$k \sum_{n=1}^m \|\partial_t \xi_n\|_0^2 \leq Ch^{2(r+q)} \|\dot{u}\|^2_{L_\infty(0,T;H^{r+1}(\Omega))}, \tag{6.13}$$

$$k \sum_{n=1}^m \|\partial_t \theta_n\|_0^2 \leq Ch^{2(r+q)} \|\dot{\sigma}\|^2_{L_\infty(0,T;H^{r+1}(\Omega))}, \tag{6.14}$$

$$k \sum_{n=1}^m \|\Delta_n u\|_0^2 \leq Ck^4 \|\ddot{u}\|^2_{L_2(0,T;L_2(\Omega))}, \tag{6.15}$$

$$k \sum_{n=1}^m \|\Delta_n \sigma\|_0^2 \leq Ck^4 \|\ddot{\sigma}\|^2_{L_2(0,T;L_2(\Omega))}. \tag{6.16}$$

Proof. Like in [RS06], use Lemma 6.1 and Taylor estimates from Lemma 5.1. As the proofs for the bounds involving σ, η_n and θ_n are directly analogous to those involving u, χ_n and ξ_n, we only need to prove the latter ones.

Firstly, (6.3) and (6.5) are immediate consequences of Lemma 6.1.

Secondly, since $\sum_{n=1}^{m} k \le T$, which is constant, we obtain (6.7) from (6.3):

$$\begin{aligned} k\sum_{n=1}^{m}\|\bar{\xi}_n\|_V^2 &\le k\sum_{n=1}^{m} Ch^{2r}\|u\|^2_{L_\infty(t_{n-1},t_n;H^{r+1}(\Omega))} \\ &\le Ch^{2r}\|u\|^2_{L_\infty(0,T;H^{r+1}(\Omega))}. \end{aligned}$$

With the same argument (6.9) follows from (6.5).

For proving (6.11) observe that, since u_0^h is the L_2-projection of $u_0 = \breve{u}$ (4.3c),

$$(u_0^h, v)_0 = (u_0, v)_0 \quad \forall\, v \in V^h.$$

This implies

$$\|\chi_0\|_0^2 = (\chi_0,\chi_0)_0 = (u_0^h - u^*, \chi_0)_0 = (u_0 - u^*, \chi_0)_0 = (\xi_0, \chi_0)_0 \le \|\xi_0\|_0 \cdot \|\chi_0\|_0.$$

Hence $\|\chi_0\|_0 \le \|\xi_0\|_0$ and combining this with (6.5) gives (6.11).

In order to prove (6.13), we have from (5.7) $\|\partial_t \xi_n\|_0 \le \|\dot{\xi}\|_{L_\infty(0,T;L_2(\Omega))}$ and by Lemma 6.1 $\|\dot{\xi}(t)\|_0 \le Ch^{r+q}\|\dot{u}(t)\|_{H^{r+1}(\Omega)}$, which can be combined to

$$\|\partial_t \xi_n\|_0 \le Ch^{r+q}\|\dot{u}\|_{L_\infty(0,T;H^{r+1}(\Omega))}.$$

This directly implies (6.13) by again using $\sum_{n=1}^{m} k \le T$.

For proving (6.15), we use (5.11):

$$k\sum_{n=1}^{m}\|\Delta_n u\|_0^2 \le k\sum_{n=1}^{m} Ck^3\|\ddot{u}\|^2_{L_2(t_{n-1},t_n;L_2(\Omega))} \le Ck^4\|\ddot{u}\|^2_{L_2(0,T;L_2(\Omega))}. \qquad \square$$

6.2 Error inequalities

In this section we establish two inequalities which will form the basis for the error estimate. One inequality is derived by using the inverse Laplacian and the other by using the $u+\sigma$ approach. In both cases we do not yet examine the term that contains the nonlinearity and leave it as it is until the next section.

Lemma 6.3 (Error inequalities).
Suppose $\mathcal{ER}^q$ holds, that $\breve{u}, \breve{\sigma} \in H^{r+1}(\Omega)$ and

$$u, \sigma \in W_\infty^1(0,T;H^{r+1}(\Omega)) \cap H^3(0,T;L_2(\Omega)).$$

Then, for any $\epsilon > 0$ and for any $m \in \{1,\dots,N\}$,

$$\begin{aligned} &\|\mathcal{G}^h\chi_m\|_V^2 + \|\eta_m\|_0^2 + k\sum_{n=1}^{m}\|\bar{\chi}_n\|_0^2 + 2k\sum_{n=1}^{m}(\mathscr{B}_n^Q - \overline{\gamma(u)\sigma}_n, \bar{\eta}_n)_0 \\ &\quad\le C\Big(1+\frac{1}{\epsilon}\Big)(k^4 + h^{2(r+q)}) + \epsilon k\sum_{n=1}^{m}\|\bar{\eta}_n\|_0^2. \end{aligned} \tag{6.17}$$

Furthermore, for all $m \in \{1,\dots,N\}$,

$$\begin{aligned} &\|\chi_m + \eta_m\|_0^2 + k\sum_{n=1}^{m}\|\bar{\chi}_n + \bar{\eta}_n\|_V^2 + 2k\sum_{n=1}^{m}(\mathscr{B}_n^Q - \overline{\gamma(u)\sigma}_n, \bar{\chi}_n + \bar{\eta}_n)_0 \\ &\quad\le C(k^4 + h^{2(r+q)}) + 6C_V^2 k\sum_{n=1}^{m}\|\bar{\chi}_n\|_0^2. \end{aligned} \tag{6.18}$$

Proof. Subtract the average of (4.3a) between t_n and t_{n-1} from (5.15a) and obtain

$$(\partial_t u_n^h, v)_0 - (\bar{\dot{u}}_n, v)_0 + (\bar{u}_n^h, v)_V - (\bar{u}_n, v)_V + (\bar{\sigma}_n^h, v)_V - (\bar{\sigma}_n, v)_V = 0 \quad \forall v \in V^h,$$

which rearranges to

$$(-\Delta_n u, v)_0 + (\partial_t \chi_n, v)_0 - (\partial_t \xi_n, v)_0 + (\bar{\chi}_n, v)_V - (\bar{\xi}_n, v)_V + (\bar{\eta}_n, v)_V - (\bar{\theta}_n, v)_V = 0 \quad \forall v \in V^h. \tag{6.19}$$

Here we have used that

$$\begin{aligned}\partial_t u_n^h - \bar{\dot{u}}_n &= \partial_t u_n^h - \partial_t u_n - \Delta_n u = \partial_t(u_n^h - u_n^*) - \partial_t(u_n - u_n^*) - \Delta_n u \\ &= \partial_t \chi_n - \partial_t \xi_n - \Delta_n u.\end{aligned}$$

Similarly, subtract the average of (4.4) between t_n and t_{n-1} from (5.15b) and obtain

$$(\partial_t \sigma_n^h, v)_0 - (\bar{\dot{\sigma}}_n, v)_0 + (\mathscr{B}_n^Q, v)_0 - (\overline{\gamma(u)\sigma}_n, v)_0 = (\bar{u}_n^h, v)_0 - (\bar{u}_n, v)_0 \quad \forall v \in V^h,$$

which rearranges to

$$(-\Delta_n \sigma, v)_0 + (\partial_t \eta_n, v)_0 - (\partial_t \theta_n, v)_0 + (\mathscr{B}_n^Q - \overline{\gamma(u)\sigma}_n, v)_0 = (\bar{\chi}_n, v)_0 - (\bar{\xi}_n, v)_0 \quad \forall v \in V^h. \tag{6.20}$$

Observe that, by the elliptic projection, $(\xi_n, v)_V = 0$ and $(\theta_n, v)_V = 0$ for any $v \in V^h$ and each n. Thus (6.19) and (6.20) give us the following system of error equations:

$$(\partial_t \chi_n, v)_0 = (\partial_t \xi_n, v)_0 + (\Delta_n u, v)_0 - (\bar{\chi}_n, v)_V - (\bar{\eta}_n, v)_V \quad \forall v \in V^h, \tag{6.21a}$$

$$(\partial_t \eta_n, v)_0 = (\partial_t \theta_n, v)_0 + (\Delta_n \sigma, v)_0 + (\bar{\chi}_n, v)_0 - (\bar{\xi}_n, v)_0 - (\mathscr{B}_n^Q - \overline{\gamma(u)\sigma}_n, v)_0 \quad \forall v \in V^h. \tag{6.21b}$$

In order to prove (6.17), choose $v = \mathcal{G}^h \bar{\chi}_n$ in (6.21a) and $v = \bar{\eta}_n$ in (6.21b), which gives us, using the property of the inverse Laplacian,

$$(\partial_t \chi_n, \mathcal{G}^h \bar{\chi}_n)_0 = (\partial_t \xi_n, \mathcal{G}^h \bar{\chi}_n)_0 + (\Delta_n u, \mathcal{G}^h \bar{\chi}_n)_0 - (\bar{\chi}_n, \bar{\chi}_n)_0 - (\bar{\eta}_n, \bar{\chi}_n)_0, \tag{6.22a}$$

$$(\partial_t \eta_n, \bar{\eta}_n)_0 = (\partial_t \theta_n, \bar{\eta}_n)_0 + (\Delta_n \sigma, \bar{\eta}_n)_0 + (\bar{\chi}_n, \bar{\eta}_n)_0 - (\bar{\xi}_n, \bar{\eta}_n)_0 - (\mathscr{B}_n^Q - \overline{\gamma(u)\sigma}_n, \bar{\eta}_n)_0. \tag{6.22b}$$

Adding these two equations gives, using (5.14),

$$\begin{aligned}&\frac{1}{2k}\big(\|\mathcal{G}^h \chi_n\|_V^2 - \|\mathcal{G}^h \chi_{n-1}\|_V^2 + \|\eta_n\|_0^2 - \|\eta_{n-1}\|_0^2\big) + \|\bar{\chi}_n\|_0^2 + (\mathscr{B}_n^Q - \overline{\gamma(u)\sigma}_n, \bar{\eta}_n)_0 \\ &\quad = (\partial_t \xi_n, \mathcal{G}^h \bar{\chi}_n)_0 + (\partial_t \theta_n, \bar{\eta}_n)_0 + (\Delta_n u, \mathcal{G}^h \bar{\chi}_n)_0 + (\Delta_n \sigma, \bar{\eta}_n)_0 - (\bar{\xi}_n, \bar{\eta}_n)_0.\end{aligned}$$

Multiplying by $2k$ and taking the sum for $n = 1, \dots, m$ gives

$$\begin{aligned}&\|\mathcal{G}^h \chi_m\|_V^2 + \|\eta_m\|_0^2 + 2k \sum_{n=1}^m \|\bar{\chi}_n\|_0^2 + 2k \sum_{n=1}^m (\mathscr{B}_n^Q - \overline{\gamma(u)\sigma}_n, \bar{\eta}_n)_0 \\ &\quad = \|\mathcal{G}^h \chi_0\|_V^2 + \|\eta_0\|_0^2 + 2k \sum_{n=1}^m (\partial_t \xi_n, \mathcal{G}^h \bar{\chi}_n)_0 + 2k \sum_{n=1}^m (\partial_t \theta_n, \bar{\eta}_n)_0 \\ &\qquad + 2k \sum_{n=1}^m (\Delta_n u, \mathcal{G}^h \bar{\chi}_n)_0 + 2k \sum_{n=1}^m (\Delta_n \sigma, \bar{\eta}_n)_0 - 2k \sum_{n=1}^m (\bar{\xi}_n, \bar{\eta}_n)_0.\end{aligned}$$

Applying the Cauchy-Schwarz inequality and Young's inequality gives

$$\begin{aligned}
&\|\mathcal{G}^h\chi_m\|_V^2 + \|\eta_m\|_0^2 + 2k\sum_{n=1}^m\|\bar\chi_n\|_0^2 + 2k\sum_{n=1}^m(\mathscr{B}_n^Q - \overline{\gamma(u)\sigma}_n, \bar\eta_n)_0 \\
&\le \|\mathcal{G}^h\chi_0\|_V^2 + \|\eta_0\|_0^2 + \frac{1}{\epsilon_1}k\sum_{n=1}^m\|\partial_t\xi_n\|_0^2 + \frac{1}{\epsilon_2}k\sum_{n=1}^m\|\partial_t\theta_n\|_0^2 \\
&\quad + \frac{1}{\epsilon_3}k\sum_{n=1}^m\|\Delta_n u\|_0^2 + \frac{1}{\epsilon_4}k\sum_{n=1}^m\|\Delta_n\sigma\|_0^2 + \frac{1}{\epsilon_5}k\sum_{n=1}^m\|\bar\xi_n\|_0^2 \\
&\quad + (\epsilon_1+\epsilon_3)k\sum_{n=1}^m\|\mathcal{G}^h\bar\chi_n\|_0^2 + (\epsilon_2+\epsilon_4+\epsilon_5)k\sum_{n=1}^m\|\bar\eta_n\|_0^2.
\end{aligned}$$

Since by (2.12), (5.13) and (2.14) $\|\mathcal{G}^h\bar\chi_n\|_0 \le C_V\|\mathcal{G}^h\bar\chi_n\|_V \le C_V\|\bar\chi_n\|_{V'} \le C_V^2\|\bar\chi_n\|_0$, we get by choosing $\epsilon_1 = \epsilon_3 = 1/(2C_V^4)$ and $\epsilon_2 = \epsilon_4 = \epsilon_5 = \epsilon/3$, where $\epsilon > 0$ is arbitrary,

$$\begin{aligned}
&\|\mathcal{G}^h\chi_m\|_V^2 + \|\eta_m\|_0^2 + k\sum_{n=1}^m\|\bar\chi_n\|_0^2 + 2k\sum_{n=1}^m(\mathscr{B}_n^Q - \overline{\gamma(u)\sigma}_n, \bar\eta_n)_0 \\
&\le \|\mathcal{G}^h\chi_0\|_V^2 + \|\eta_0\|_0^2 + 2C_V^4k\sum_{n=1}^m\|\partial_t\xi_n\|_0^2 + \frac{3}{\epsilon}k\sum_{n=1}^m\|\partial_t\theta_n\|_0^2 \\
&\quad + 2C_V^4k\sum_{n=1}^m\|\Delta_n u\|_0^2 + \frac{3}{\epsilon}k\sum_{n=1}^m\|\Delta_n\sigma\|_0^2 + \frac{3}{\epsilon}k\sum_{n=1}^m\|\bar\xi_n\|_0^2 \\
&\quad + \epsilon k\sum_{n=1}^m\|\bar\eta_n\|_0^2.
\end{aligned} \tag{6.23}$$

By (5.13), $\|\mathcal{G}^h\chi_0\|_V \le C_V\|\chi_0\|_0$ and applying Lemma 6.2 yields (6.17).

To prove the second inequality (6.18), add (6.21a) and (6.21b) and choose $v = \bar\chi_n + \bar\eta_n$, which yields

$$\begin{aligned}
&(\partial_t(\chi_n+\eta_n), \bar\chi_n+\bar\eta_n)_0 + (\bar\chi_n+\bar\eta_n, \bar\chi_n+\bar\eta_n)_V + (\mathscr{B}_n^Q - \overline{\gamma(u)\sigma}_n, \bar\chi_n+\bar\eta_n)_0 \\
&\quad = (\partial_t\xi_n, \bar\chi_n+\bar\eta_n)_0 + (\partial_t\theta_n, \bar\chi_n+\bar\eta_n)_0 + (\Delta_n u, \bar\chi_n+\bar\eta_n)_0 + (\Delta_n\sigma, \bar\chi_n+\bar\eta_n)_0 \\
&\qquad + (\bar\chi_n, \bar\chi_n+\bar\eta_n)_0 - (\bar\xi_n, \bar\chi_n+\bar\eta_n)_0.
\end{aligned}$$

Using Cauchy-Schwarz inequality and $\|v\|_0 \le C_V\|v\|_V$ and Young's inequality, we obtain

$$\begin{aligned}
&\frac{1}{2k}\|\chi_n+\eta_n\|_0^2 - \frac{1}{2k}\|\chi_{n-1}+\eta_{n-1}\|_0^2 + \|\bar\chi_n+\bar\eta_n\|_V^2 + (\mathscr{B}_n^Q - \overline{\gamma(u)\sigma}_n, \bar\chi_n+\bar\eta_n)_0 \\
&\quad \le \frac{C_V^2}{2\epsilon}\|\partial_t\xi_n\|_0^2 + \frac{C_V^2}{2\epsilon}\|\partial_t\theta_n\|_0^2 + \frac{C_V^2}{2\epsilon}\|\Delta_n u\|_0^2 + \frac{C_V^2}{2\epsilon}\|\Delta_n\sigma\|_0^2 \\
&\qquad + \frac{C_V^2}{2\epsilon}\|\bar\chi_n\|_0^2 + \frac{C_V^2}{2\epsilon}\|\bar\xi_n\|_0^2 + \frac{6\epsilon}{2}\|\bar\chi_n+\bar\eta_n\|_V^2.
\end{aligned}$$

We choose $\epsilon = 1/6$ to cancel the term $\|\bar\chi_n+\bar\eta_n\|_V^2$ on the right-hand side:

$$\begin{aligned}
&\frac{1}{2k}\|\chi_n+\eta_n\|_0^2 - \frac{1}{2k}\|\chi_{n-1}+\eta_{n-1}\|_0^2 + \frac{1}{2}\|\bar\chi_n+\bar\eta_n\|_V^2 + (\mathscr{B}_n^Q - \overline{\gamma(u)\sigma}_n, \bar\chi_n+\bar\eta_n)_0 \\
&\quad \le 3C_V^2(\|\partial_t\xi_n\|_0^2 + \|\partial_t\theta_n\|_0^2 + \|\Delta_n u\|_0^2 + \|\Delta_n\sigma\|_0^2 + \|\bar\chi_n\|_0^2 + \|\bar\xi_n\|_0^2).
\end{aligned} \tag{6.24}$$

Multiplying by $2k$ and taking the sum for $n = 1, \dots, m$ yields

$$\begin{aligned}\|\chi_m + \eta_m\|_0^2 - \|\chi_0 + \eta_0\|_0^2 + k\sum_{n=1}^{m}\|\bar{\chi}_n + \bar{\eta}_n\|_V^2 + 2k\sum_{n=1}^{m}(\mathscr{B}_n^Q - \overline{\gamma(u)\sigma}_n, \bar{\chi}_n + \bar{\eta}_n)_0 \\ \leq 6C_V^2 k\sum_{n=1}^{m}\left(\|\partial_t\xi_n\|_0^2 + \|\partial_t\theta_n\|_0^2 + \|\Delta_n u\|_0^2 + \|\Delta_n\sigma\|_0^2 + \|\bar{\chi}_n\|_0^2 + \|\bar{\xi}_n\|_0^2\right).\end{aligned} \tag{6.25}$$

Now we use the bounds for the terms involving ξ_n, θ_n, $\Delta_n u$, $\Delta_n\sigma$, χ_0 and η_0 from Lemma 6.2 and thus obtain

$$\begin{aligned}\|\chi_m + \eta_m\|_0^2 + k\sum_{n=1}^{m}\|\bar{\chi}_n + \bar{\eta}_n\|_V^2 + 2k\sum_{n=1}^{m}(\mathscr{B}_n^Q - \overline{\gamma(u)\sigma}_n, \bar{\chi}_n + \bar{\eta}_n)_0 \\ \leq C(k^4 + h^{2(r+q)}) + 6C_V^2 k\sum_{n=1}^{m}\|\bar{\chi}_n\|_0^2,\end{aligned} \tag{6.26}$$

which is the second inequality we wanted to prove. □

6.3 Nonlinearity errors

The error inequalities which were derived in Lemma 6.3 in the previous section contain terms with the nonlinearity and its approximation. These terms have the following form

$$2k\sum_{n=1}^{m}(\mathscr{B}_n^Q - \overline{\gamma(u)\sigma}_n, \bar{v}_n)_0 \tag{6.27}$$

where $v_n = \eta_n$ in the case of (6.17) or $v_n = \chi_n + \eta_n$ in the case of (6.18). In this section we give bounds for these terms and have to distinguish between the various methods.

Since the terms (6.27) appear on the left-hand side of (6.17) and (6.18) we bound them from below instead of taking the absolute value and bounding them from above. Another reason for bounding below is that (6.27) contains helpful terms, as we will see later.

This section contains the following results:

- Lemma 6.4 gives a bound for the term (6.27) for methods $Q \in \{2, 5\}$, going as far as possible without distinguishing between $Q = 2$ and $Q = 5$.
- Lemma 6.5 does the same but for methods $Q \in \{1, 4\}$.
- Lemma 6.6 has bounds for the remaining terms from the previous lemmas. These terms contain the extrapolation errors. As opposed to the previous lemmas, here, the linearised methods $Q \in \{1, 2\}$ can be dealt with simultaneously, as can the nonlinear methods $Q \in \{4, 5\}$ (which do not use extrapolation).
- Lemma 6.7 combines the previous three lemmas and presents a unified nonlinearity error. This result will be used in Section 6.4.

For the next three lemmas, we make no specific choice for v_n in (6.27) and just assume that $v_n \in L_2(\Omega)$ for $n = 0, \dots, N$. We will make the choice of either $v_n = \eta_n$ or $v_n = \chi_n + \eta_n$ only in Lemma 6.7.

Lemma 6.4 (Intermediate nonlinearity error for methods $Q \in \{2,5\}$). *Suppose that $\mathcal{ER}^q$ holds and that*

$$u \in W^1_\infty(0,T;L_4(\Omega)) \text{ and}$$
$$\sigma \in L_\infty(0,T;H^{r+1}(\Omega)) \cap W^1_\infty(0,T;L_4(\Omega)).$$

Let $v_n \in L_2(\Omega)$ for $n = 0,\dots,N$ and the method $Q \in \{2,5\}$. Then for any $\epsilon > 0$ and any $m \in \{1,\dots,N\}$,

$$\begin{aligned} 2k\sum_{n=1}^{m}(\mathscr{B}_n^Q - \overline{\gamma(u)\sigma}_n, \bar{v}_n)_0 \geq\ & 2k\sum_{n=1}^{m}(\gamma(\mathscr{E}_n^Q u^h)\bar{\eta}_n, \bar{v}_n)_0 \\ & - \epsilon k\sum_{n=1}^{m}\|\bar{v}_n\|_0^2 - \frac{C}{\epsilon}(k^4 + h^{2(r+q)}) \\ & - 2c_1 k\sum_{n=1}^{m}\|\mathscr{E}_n^Q u^h - \bar{u}_n\|_0 \cdot \|\bar{v}_n\|_0, \end{aligned} \tag{6.28}$$

where $c_1 = C'_\gamma \|\sigma\|_{L_\infty(0,T;L_\infty(\Omega))}$.

Proof. For $Q \in \{2,5\}$, setting $\tilde{\gamma}_n = \gamma(\mathscr{E}_n^Q u^h)$,

$$\begin{aligned} \mathscr{B}_n^Q - \overline{\gamma(u)\sigma}_n &= \tilde{\gamma}_n\bar{\sigma}_n^h - \overline{\gamma(u)\sigma}_n \\ &= (\tilde{\gamma}_n\bar{\sigma}_n^h - \tilde{\gamma}_n\bar{\sigma}_n) + (\tilde{\gamma}_n\bar{\sigma}_n - \overline{\gamma(u)}_n\bar{\sigma}_n) + \big(\overline{\gamma(u)}_n\bar{\sigma}_n - \overline{\gamma(u)\sigma}_n\big) \\ &= \tilde{\gamma}_n\bar{\eta}_n - \tilde{\gamma}_n\bar{\theta}_n + \big(\tilde{\gamma}_n - \overline{\gamma(u)}_n\big)\,\bar{\sigma}_n - \tfrac{1}{4}\big(\gamma(u_n) - \gamma(u_{n-1})\big)(\sigma_n - \sigma_{n-1}), \end{aligned}$$

where we have used the fact that for any real numbers a, b, c, d,

$$\tfrac{1}{2}(a+b)\cdot\tfrac{1}{2}(c+d) - \tfrac{1}{2}(ac+bd) = -\tfrac{1}{4}(a-b)(c-d).$$

Thus, if we put, for convenience of notation,

$$A := \tilde{\gamma}_n\bar{\theta}_n - (\tilde{\gamma}_n - \overline{\gamma(u)}_n)\,\bar{\sigma}_n + \tfrac{1}{4}(\gamma(u_n) - \gamma(u_{n-1}))(\sigma_n - \sigma_{n-1}),$$

then

$$\mathscr{B}_n^Q - \overline{\gamma(u)\sigma}_n = \tilde{\gamma}_n\bar{\eta}_n - A,$$

and taking the inner product with $\bar{v}_n$, summing over n and multiplying by $2k$ gives us,

$$\begin{aligned} 2k\sum_{n=1}^{m}(\mathscr{B}_n^Q - \overline{\gamma(u)\sigma}_n, \bar{v}_n)_0 &= 2k\sum_{n=1}^{m}(\tilde{\gamma}_n\bar{\eta}_n, \bar{v}_n)_0 - 2k\sum_{n=1}^{m}(A, \bar{v}_n)_0 \\ &\geq 2k\sum_{n=1}^{m}(\tilde{\gamma}_n\bar{\eta}_n, \bar{v}_n)_0 - 2k\sum_{n=1}^{m}\|A\|_0 \cdot \|\bar{v}_n\|_0. \end{aligned} \tag{6.29}$$

Looking specifically at the norm of A in the last inequality, we get

$$\begin{aligned} \|A\|_0 \leq\ & \hat{\gamma}\|\bar{\theta}_n\|_0 + \|\tilde{\gamma}_n - \overline{\gamma(u)}_n\|_0 \cdot \|\bar{\sigma}_n\|_{L_\infty(\Omega)} \\ & + \tfrac{1}{4}\|\gamma(u_n) - \gamma(u_{n-1})\|_{L_4(\Omega)} \cdot \|\sigma_n - \sigma_{n-1}\|_{L_4(\Omega)}. \end{aligned} \tag{6.30}$$

Now, we find bounds for each term in (6.30). To start with, observe that

$$\|\bar{\sigma}_n\|_{L_\infty(\Omega)} \leq \|\sigma\|_{L_\infty(0,T;L_\infty(\Omega))}.$$

Next, by (5.7),

$$\|\sigma_n - \sigma_{n-1}\|_{L_4(\Omega)} \le k\|\dot\sigma\|_{L_\infty(t_{n-1},t_n;L_4(\Omega))}.$$

Similarly, using Lemma 3.9 and (5.7),

$$\|\gamma(u_n) - \gamma(u_{n-1})\|_{L_4(\Omega)} \le C'_\gamma\, k\|\dot u\|_{L_\infty(t_{n-1},t_n;L_4(\Omega))}.$$

Also, using Lemmas 3.9 and 3.11 and again (5.7)

$$\begin{aligned}\|\gamma(\mathscr{E}_n^Q u^h) - \overline{\gamma(u)}_n\|_0 &\le \|\gamma(\mathscr{E}_n^Q u^h) - \gamma(\bar u_n)\|_0 + \|\gamma(\bar u_n) - \overline{\gamma(u)}_n\|_0\\ &\le C'_\gamma\,\|\mathscr{E}_n^Q u^h - \bar u_n\|_0 + \frac{C''_\gamma}{8}\|u_n - u_{n-1}\|^2_{L_4(\Omega)}\\ &\le C'_\gamma\,\|\mathscr{E}_n^Q u^h - \bar u_n\|_0 + \frac{C''_\gamma}{8}k^2\|\dot u\|^2_{L_\infty(t_{n-1},t_n;L_4(\Omega))}.\end{aligned}$$

Combining all these and inserting them in (6.30), we get

$$\|A\|_0 \le \hat\gamma\|\bar\theta_n\|_0 + c_1\|\mathscr{E}_n^Q u^h - \bar u_n\|_0 + c_2\,k^2, \tag{6.31}$$

where (having adapted the norms: $\|v\|_{L_\infty(t_{n-1},t_n;X)} \le \|v\|_{L_\infty(0,T;X)}$)

$$\begin{aligned}c_1 &= C'_\gamma\,\|\sigma\|_{L_\infty(0,T;L_\infty(\Omega))},\\ c_2 &= \frac{C''_\gamma}{8}\,\|\dot u\|_{L_\infty(0,T;L_4(\Omega))}\cdot\|\sigma\|_{L_\infty(0,T;L_\infty(\Omega))}\\ &\quad + \frac{C'_\gamma}{4}\,\|\dot u\|_{L_\infty(0,T;L_4(\Omega))}\cdot\|\dot\sigma\|_{L_\infty(0,T;L_4(\Omega))}.\end{aligned}$$

Inserting (6.31) in (6.29) gives

$$\begin{aligned}&2k\sum_{n=1}^m(\mathscr{B}_n^Q - \overline{\gamma(u)\sigma}_n, \bar v_n)_0\\ &\quad\ge 2k\sum_{n=1}^m(\tilde\gamma_n\bar\eta_n, \bar v_n)_0 - 2k\sum_{n=1}^m\big(\hat\gamma\|\bar\theta_n\|_0 + c_1\|\mathscr{E}_n^Q u^h - \bar u_n\|_0 + c_2\,k^2\big)\cdot\|\bar v_n\|_0.\end{aligned}$$

Using Young's inequality we obtain,

$$\begin{aligned}2k\sum_{n=1}^m(\mathscr{B}_n^Q - \overline{\gamma(u)\sigma}_n, \bar v_n)_0 \ge 2k\sum_{n=1}^m(\tilde\gamma_n\bar\eta_n, \bar v_n)_0 &- \frac{1}{\epsilon}k\sum_{n=1}^m(\hat\gamma\|\bar\theta_n\|_0 + c_2k^2)^2\\ &- \epsilon k\sum_{n=1}^m\|\bar v_n\|_0^2 - 2c_1k\sum_{n=1}^m\|\mathscr{E}_n^Q u^h - \bar u_n\|_0\cdot\|\bar v_n\|_0.\end{aligned}$$

Rearranging this and using $\sum_{n=1}^m k = T$, finally gives us

$$\begin{aligned}2k\sum_{n=1}^m(\mathscr{B}_n^Q - \overline{\gamma(u)\sigma}_n, \bar v_n)_0 \ge 2k\sum_{n=1}^m(\tilde\gamma_n\bar\eta_n, \bar v_n)_0 &- \frac{2\hat\gamma^2}{\epsilon}k\sum_{n=1}^m\|\bar\theta_n\|_0^2 - \frac{2c_2^2T}{\epsilon}k^4\\ &- \epsilon k\sum_{n=1}^m\|\bar v_n\|_0^2 - 2c_1k\sum_{n=1}^m\|\mathscr{E}_n^Q u^h - \bar u_n\|_0\cdot\|\bar v_n\|_0.\end{aligned}$$

An application of (6.10) from Lemma 6.2 now yields inequality (6.28). To conclude we read off the regularity requirements on u and σ from the norms appearing in (6.8) and in the definitions of c_1 and c_2. Note that $\sigma \in L_\infty(0,T;H^{r+1}(\Omega))$ implies $\sigma \in L_\infty(0,T;L_\infty(\Omega))$, since $r \ge 1$ and $d \le 3$. □

Lemma 6.5 (Intermediate nonlinearity error for methods $Q \in \{1,4\}$). *Suppose $\mathcal{ER}^q$ holds, that $\breve{u}, \breve{\sigma} \in H^{r+1}(\Omega)$ and that*

$$u, \sigma \in L_\infty(0,T; H^{r+1}(\Omega)).$$

Let also $v_n \in L_2(\Omega)$ for $n = 0, \dots, N$ and the method $Q \in \{1,4\}$. Then for any $\epsilon > 0$ and for any $m \in \{1, \dots, N\}$,

$$\begin{aligned} 2k\sum_{n=1}^{m}(\mathscr{B}_n^Q - \overline{\gamma(u)\sigma}_n, \bar{v}_n)_0 \geq 2k\sum_{n=1}^{m}(\gamma(\mathscr{E}_n^Q u^h)\bar{\eta}_n, \bar{v}_n)_0 - \epsilon k\sum_{n=1}^{m}\|\bar{v}_n\|_0^2 \\ - \frac{C}{\epsilon}\,h^{2(r+q)} - \frac{2c_1^2}{\epsilon}k\sum_{n=1}^{m-1}\|\chi_n\|_0^2 - \frac{4\hat{\gamma}^2}{\epsilon}k\sum_{n=1}^{m-1}\|\eta_n\|_0^2 \\ - c_1 k\sum_{n=1}^{m}\|\mathscr{E}_n^Q u^h - u_n\|_0 \cdot \|\bar{v}_n\|_0, \end{aligned} \tag{6.32}$$

where $c_1 = C'_\gamma \|\sigma\|_{L_\infty(0,T;L_\infty(\Omega))}$.

Proof. For $Q \in \{1,4\}$ we have, setting $\tilde{\gamma}_n := \gamma(\mathscr{E}_n^Q u^h)$,

$$\begin{aligned} \mathscr{B}_n^Q - \overline{\gamma(u)\sigma}_n &= \tfrac{1}{2}\tilde{\gamma}_n\sigma_n^h + \tfrac{1}{2}\gamma(u_{n-1}^h)\sigma_{n-1}^h - \tfrac{1}{2}\gamma(u_n)\sigma_n - \tfrac{1}{2}\gamma(u_{n-1})\sigma_{n-1} \\ &= \tfrac{1}{2}\tilde{\gamma}_n(\sigma_n^h - \sigma_n) + \tfrac{1}{2}\gamma(u_{n-1}^h)(\sigma_{n-1}^h - \sigma_{n-1}) \\ &\quad + \tfrac{1}{2}\big(\tilde{\gamma}_n - \gamma(u_n)\big)\sigma_n + \tfrac{1}{2}\big(\gamma(u_{n-1}^h) - \gamma(u_{n-1})\big)\sigma_{n-1}. \end{aligned} \tag{6.33}$$

We can further rearrange

$$\begin{aligned} &\tfrac{1}{2}\tilde{\gamma}_n(\sigma_n^h - \sigma_n) + \tfrac{1}{2}\gamma(u_{n-1}^h)(\sigma_{n-1}^h - \sigma_{n-1}) \\ &\quad = \tilde{\gamma}_n(\bar{\sigma}_n^h - \bar{\sigma}_n) - \tfrac{1}{2}\big(\tilde{\gamma}_n - \gamma(u_{n-1}^h)\big)(\sigma_{n-1}^h - \sigma_{n-1}) \\ &\quad = \tilde{\gamma}_n(\bar{\eta}_n - \bar{\theta}_n) - \tfrac{1}{2}\big(\tilde{\gamma}_n - \gamma(u_{n-1}^h)\big)(\eta_{n-1} - \theta_{n-1}). \end{aligned}$$

Inserting this in (6.33) gives

$$\begin{aligned} \mathscr{B}_n^Q - \overline{\gamma(u)\sigma}_n = \tilde{\gamma}_n(\bar{\eta}_n - \bar{\theta}_n) - \tfrac{1}{2}\big(\tilde{\gamma}_n - \gamma(u_{n-1}^h)\big)(\eta_{n-1} - \theta_{n-1}) \\ + \tfrac{1}{2}\big(\tilde{\gamma}_n - \gamma(u_n)\big)\sigma_n + \tfrac{1}{2}\big(\gamma(u_{n-1}^h) - \gamma(u_{n-1})\big)\sigma_{n-1}. \end{aligned}$$

Thus, taking the inner product with $\bar{v}_n$, multiplying by $2k$ and summing up this equation for $n = 1, \dots, N$ gives,

$$\begin{aligned} 2k\sum_{n=1}^{m}(\mathscr{B}_n^Q - \overline{\gamma(u)\sigma}_n, \bar{v}_n)_0 = 2k\sum_{n=1}^{m}(\tilde{\gamma}_n\bar{\eta}_n, \bar{v}_n)_0 - 2k\sum_{n=1}^{m}\big(\tilde{\gamma}_n\bar{\theta}_n, \bar{v}_n\big)_0 \\ - k\sum_{n=1}^{m}\big((\tilde{\gamma}_n - \gamma(u_{n-1}^h))(\eta_{n-1} - \theta_{n-1}), \bar{v}_n\big)_0 \\ + k\sum_{n=1}^{m}\big((\tilde{\gamma}_n - \gamma(u_n))\sigma_n + (\gamma(u_{n-1}^h) - \gamma(u_{n-1}))\sigma_{n-1}, \bar{v}_n\big)_0. \end{aligned}$$

Using the Cauchy-Schwarz inequality,

$$\begin{aligned} 2k\sum_{n=1}^{m}(\mathscr{B}_n^Q - \overline{\gamma(u)\sigma}_n, \bar{v}_n)_0 \geq 2k\sum_{n=1}^{m}(\tilde{\gamma}_n\bar{\eta}_n, \bar{v}_n)_0 - 2k\sum_{n=1}^{m}\|\tilde{\gamma}_n\bar{\theta}_n\|_0 \cdot \|\bar{v}_n\|_0 \\ - k\sum_{n=1}^{m}\|(\tilde{\gamma}_n - \gamma(u_{n-1}^h))(\eta_{n-1} - \theta_{n-1})\|_0 \cdot \|\bar{v}_n\|_0 \\ - k\sum_{n=1}^{m}\|(\tilde{\gamma}_n - \gamma(u_n))\sigma_n + (\gamma(u_{n-1}^h) - \gamma(u_{n-1}))\sigma_{n-1}\|_0 \cdot \|\bar{v}_n\|_0. \end{aligned} \tag{6.34}$$

We bound some of the terms in this inequality now. Firstly,

$$\begin{aligned}\|(\tilde{\gamma}_n - \gamma(u^h_{n-1}))(\eta_{n-1} - \theta_{n-1})\|_0 &\le \|\tilde{\gamma}_n - \gamma(u^h_{n-1})\|_{L_\infty(\Omega)} \cdot \|\eta_{n-1} - \theta_{n-1}\|_0 \\ &\le 2\hat{\gamma}\|\eta_{n-1} - \theta_{n-1}\|_0 \\ &\le 2\hat{\gamma}(\|\eta_{n-1}\|_0 + \|\theta_{n-1}\|_0). \end{aligned} \tag{6.35}$$

Secondly,

$$\begin{aligned}&\|(\tilde{\gamma}_n - \gamma(u_n))\sigma_n + (\gamma(u^h_{n-1}) - \gamma(u_{n-1}))\sigma_{n-1}\|_0 \\ &\quad\le \|\tilde{\gamma}_n - \gamma(u_n)\|_0 \cdot \|\sigma_n\|_{L_\infty(\Omega)} + \|\gamma(u^h_{n-1}) - \gamma(u_{n-1})\|_0 \cdot \|\sigma_{n-1}\|_{L_\infty(\Omega)} \\ &\quad\le \|\sigma\|_{L_\infty(0,T;L_\infty(\Omega))} \left(\|\gamma(\mathscr{E}^Q_n u^h) - \gamma(u_n)\|_0 + \|\gamma(u^h_{n-1}) - \gamma(u_{n-1})\|_0\right) \\ &\quad\le C'_\gamma \|\sigma\|_{L_\infty(0,T;L_\infty(\Omega))} \left(\|\mathscr{E}^Q_n u^h - u_n\|_0 + \|u^h_{n-1} - u_{n-1}\|_0\right),\end{aligned}$$

where we have used Lemma 3.9. Defining $c_1 = C'_\gamma \|\sigma\|_{L_\infty(0,T;L_\infty(\Omega))}$, we obtain

$$\begin{aligned}&\|(\tilde{\gamma}_n - \gamma(u_n))\sigma_n + (\gamma(u^h_{n-1}) - \gamma(u_{n-1}))\sigma_{n-1}\|_0 \\ &\qquad\le c_1\|\mathscr{E}^Q_n u^h - u_n\|_0 + c_1\left(\|\chi_{n-1}\|_0 + \|\xi_{n-1}\|_0\right).\end{aligned} \tag{6.36}$$

Using (6.35) and (6.36) in (6.34) and applying Young's inequality yields

$$\begin{aligned}2k\sum_{n=1}^m(\mathscr{B}^Q_n - \overline{\gamma(u)\sigma}_n, \bar{v}_n)_0 &\ge 2k\sum_{n=1}^m(\tilde{\gamma}_n\bar{\eta}_n, \bar{v}_n)_0 - (\epsilon_1+\epsilon_2+\epsilon_3)k\sum_{n=1}^m\|\bar{v}_n\|_0^2 \\ &- \frac{\hat{\gamma}^2}{\epsilon_1}k\sum_{n=1}^m\|\bar{\theta}_n\|_0^2 - \frac{\hat{\gamma}^2}{\epsilon_2}k\sum_{n=1}^m(\|\eta_{n-1}\|_0 + \|\theta_{n-1}\|_0)^2 \\ &- \frac{c_1^2}{4\epsilon_3}k\sum_{n=1}^m(\|\chi_{n-1}\|_0 + \|\xi_{n-1}\|_0)^2 - c_1k\sum_{n=1}^m\|\mathscr{E}^Q_n u^h - u_n\|_0 \cdot \|\bar{v}_n\|_0.\end{aligned}$$

Let $\epsilon > 0$ be arbitrary and choose $\epsilon_1 = \epsilon_3 = \epsilon/4$ and $\epsilon_2 = \epsilon/2$,

$$\begin{aligned}2k\sum_{n=1}^m(\mathscr{B}^Q_n - \overline{\gamma(u)\sigma}_n, \bar{v}_n)_0 &\ge 2k\sum_{n=1}^m(\tilde{\gamma}_n\bar{\eta}_n, \bar{v}_n)_0 - \epsilon k\sum_{n=1}^m\|\bar{v}_n\|_0^2 \\ &- \frac{4\hat{\gamma}^2}{\epsilon}k\sum_{n=1}^m\|\bar{\theta}_n\|_0^2 - \frac{4\hat{\gamma}^2}{\epsilon}k\sum_{n=0}^{m-1}(\|\eta_n\|_0^2 + \|\theta_n\|_0^2) \\ &- \frac{2c_1^2}{\epsilon}k\sum_{n=0}^{m-1}(\|\chi_n\|_0^2 + \|\xi_n\|_0^2) - c_1k\sum_{n=1}^m\|\mathscr{E}^Q_n u^h - u_n\|_0 \cdot \|\bar{v}_n\|_0.\end{aligned}$$

An application of (6.9), (6.10), (6.11) and (6.12) from Lemma 6.2 finishes the proof. This last step also requires the regularity assumptions on u, σ, $\breve{u}$ and $\breve{\sigma}$. □

Lemma 6.6 (Extrapolation errors).
Suppose that $\mathcal{ER}^q$ holds, that $\breve{u}, \breve{\sigma} \in H^{r+1}(\Omega)$ and that $u \in L_\infty(0,T;H^{r+1}(\Omega))$. For methods $Q \in \{1,2\}$ we also suppose that $u \in H^2(0,T;L_2(\Omega))$. Let also $v_n \in L_2(\Omega)$ for $n = 0,\dots,N$ such that for the first element $\|v_0\|_0^2 \le Ch^{2(r+q)}$. There is a constant $C > 0$ such that for all $\epsilon, \epsilon_3, \epsilon_4 > 0$ and for all $m \in \{1,\dots,N\}$ these inequalities hold: for methods $Q = 1$ and $Q = 2$,

$$\left.\begin{aligned}k\sum_{n=1}^m\|\mathscr{E}^1_n u^h - u_n\|_0 \cdot \|\bar{v}_n\|_0 \\ 2k\sum_{n=1}^m\|\mathscr{E}^2_n u^h - \bar{u}_n\|_0 \cdot \|\bar{v}_n\|_0\end{aligned}\right\} \le C\cdot(1 + \tfrac{1}{\epsilon_3} + \tfrac{1}{\epsilon_4})\cdot(k^4 + h^{2(r+q)}) + \epsilon_4\|v_1\|_0^2 \\ + \epsilon_3 k\sum_{n=2}^m\|\bar{v}_n\|_0^2 + \frac{10}{\epsilon_3}k\sum_{n=1}^{m-1}\|\chi_n\|_0^2, \tag{6.37}$$

and for methods $Q = 4$ and $Q = 5$,

$$\left.\begin{aligned} k\sum_{n=1}^{m}\|\mathscr{E}_n^4 u^h - u_n\|_0 \cdot \|\bar{v}_n\|_0 \\ 2k\sum_{n=1}^{m}\|\mathscr{E}_n^5 u^h - \bar{u}_n\|_0 \cdot \|\bar{v}_n\|_0 \end{aligned}\right\} \leq \epsilon k\sum_{n=1}^{m}\|\bar{v}_n\|_0^2 + \frac{C}{\epsilon}h^{2(r+q)} + \frac{2}{\epsilon}k\sum_{n=1}^{m-1}\|\chi_n\|_0^2 + \frac{4}{\epsilon}k\|\bar{\chi}_m\|_0^2. \tag{6.38}$$

Proof. To have a common notation for all methods, we define for this proof,

$$E_n^Q := \begin{cases} \|\mathscr{E}_n^Q u^h - u_n\|_0 & \text{if } Q \in \{1,4\}, \\ 2\|\mathscr{E}_n^Q u^h - \bar{u}_n\|_0 & \text{if } Q \in \{2,5\}. \end{cases} \tag{6.39}$$

To start, we examine E_n^Q for each method separately. Firstly, for $Q = 1$ and $n \geq 2$, using the Taylor estimate (5.10):

$$\begin{aligned} \|\mathscr{E}_n^1 u^h - u_n\|_0 &= \|2u_{n-1}^h - u_{n-2}^h - u_n\|_0 \\ &= \|2(u_{n-1}^h - u_{n-1}) - (u_{n-2}^h - u_{n-2}) - (u_n - 2u_{n-1} + u_{n-2})\|_0 \\ &\leq 2\|u_{n-1}^h - u_{n-1}\|_0 + \|u_{n-2}^h - u_{n-2}\|_0 + \|u_n - 2u_{n-1} + u_{n-2}\|_0 \\ &\leq 2\|\chi_{n-1}\|_0 + 2\|\xi_{n-1}\|_0 + \|\chi_{n-2}\|_0 + \|\xi_{n-2}\|_0 \\ &\quad + Ck^{3/2}\|\ddot{u}\|_{L_2(t_{n-2},t_n;L_2(\Omega))}, \end{aligned} \tag{6.40}$$

while for $Q = 1$ and $n = 1$, using the Taylor estimate (5.7) and (6.11),

$$\begin{aligned} \|\mathscr{E}_1^1 u^h - u_1\|_0 = \|u_0^h - u_1\|_0 &= \|(u_0^h - u_0) - (u_1 - u_0)\|_0 \\ &= \|(\chi_0 - \xi_0) - (u_1 - u_0)\|_0 \\ &\leq \|\chi_0\|_0 + \|\xi_0\|_0 + k\|\partial_t u_1\| \\ &\leq 2\|\xi_0\|_0 + k\|\dot{u}\|_{L_\infty(0,t_1;L_2(\Omega))}. \end{aligned} \tag{6.41}$$

Secondly, for $Q = 2$ and $n \geq 2$, using the Taylor estimate (5.10),

$$\begin{aligned} \|\mathscr{E}_n^2 u^h - \bar{u}_n\|_0 &= \|\tfrac{3}{2}u_{n-1}^h - \tfrac{1}{2}u_{n-2}^h - \tfrac{1}{2}u_n - \tfrac{1}{2}u_{n-1}\|_0 \\ &= \|\tfrac{3}{2}(u_{n-1}^h - u_{n-1}) - \tfrac{1}{2}(u_{n-2}^h - u_{n-2}) - \tfrac{1}{2}(u_n - 2u_{n-1} + u_{n-2})\|_0 \\ &\leq \|\tfrac{3}{2}(\chi_{n-1} - \xi_{n-1}) - \tfrac{1}{2}(\chi_{n-2} - \xi_{n-2})\|_0 + \tfrac{1}{2}\|u_n - 2u_{n-1} + u_{n-2}\|_0 \\ &\leq \tfrac{3}{2}\|\chi_{n-1}\|_0 + \tfrac{1}{2}\|\chi_{n-2}\|_0 + \tfrac{3}{2}\|\xi_{n-1}\|_0 + \tfrac{1}{2}\|\xi_{n-2}\|_0 \\ &\quad + Ck^{3/2}\|\ddot{u}\|_{L_2(t_{n-2},t_n;L_2(\Omega))}, \end{aligned} \tag{6.42}$$

while for $Q = 2$ and $n = 1$, using the Taylor estimate (5.7) and (6.11),

$$\begin{aligned} \|\mathscr{E}_1^2 u^h - \bar{u}_1\|_0 = \|u_0^h - \bar{u}_1\|_0 &= \|u_0^h - u_0 - \tfrac{1}{2}(u_1 - u_0)\|_0 \\ &= \|\chi_0 - \xi_0 - \tfrac{1}{2}k\partial_t u_1\|_0 \\ &\leq \|\chi_0\|_0 + \|\xi_0\|_0 + \tfrac{1}{2}k\|\partial_t u_1\|_0 \\ &\leq 2\|\xi_0\|_0 + \tfrac{1}{2}k\|\dot{u}\|_{L_\infty(0,t_1;L_2(\Omega))}. \end{aligned} \tag{6.43}$$

For $Q = 4$, we obtain

$$\|\mathscr{E}_n^4 u^h - u_n\|_0 = \|u_n^h - u_n\|_0 = \|\chi_n - \xi_n\|_0 \leq \|\chi_n\|_0 + \|\xi_n\|_0. \tag{6.44}$$

Finally, for $Q = 5$, we obtain

$$\|\mathscr{E}_n^5 u^h - \bar{u}_n\|_0 = \|\bar{u}_n^h - \bar{u}_n\|_0 = \|\bar{\chi}_n - \bar{\xi}_n\|_0 \le \|\bar{\chi}_n\|_0 + \|\bar{\xi}_n\|_0. \tag{6.45}$$

In order to prove the first inequality, (6.37), we observe that for $Q = 1$ and $Q = 2$, from (6.40), (6.41), (6.42) and (6.43),

$$E_n^Q \le \begin{cases} 3\|\chi_{n-1}\|_0 + \|\chi_{n-2}\|_0 + Ch^{r+q} \\ \quad +Ck^{3/2}\|\ddot{u}\|_{L_2(t_{n-2},t_n;L_2(\Omega))} & \text{if } n \ge 2, \\ Ch^{r+q} + k\|\dot{u}\|_{L_\infty(0,t_1;L_2(\Omega))} & \text{if } n = 1, \end{cases} \tag{6.46}$$

where the bounds for $\|\xi_n\|_0$ from (6.5) have also been used. Now we form the term on the left-hand side of (6.37), split the sum, and use the fact that $\|\bar{v}_1\|_0^2 \le \frac{1}{2}(\|v_1\|_0^2 + \|v_0\|_0^2)$,

$$\begin{aligned} k\sum_{n=1}^m E_n^Q \cdot \|\bar{v}_n\|_0 &= k\sum_{n=2}^m E_n^Q \cdot \|\bar{v}_n\|_0 + kE_1^Q \cdot \|\bar{v}_1\|_0 \\ &\le k\sum_{n=2}^m E_n^Q \cdot \|\bar{v}_n\|_0 + \frac{k}{2}E_1^Q \cdot (\|v_1\|_0 + \|v_0\|_0). \end{aligned}$$

Using Young's inequality gives us

$$\begin{aligned} k\sum_{n=1}^m E_n^Q \cdot \|\bar{v}_n\|_0 \le \frac{1}{4\epsilon_3}k\sum_{n=2}^m (E_n^Q)^2 + \epsilon_3 k\sum_{n=2}^m \|\bar{v}_n\|_0^2 \\ + \Big(\frac{1}{4} + \frac{1}{4\epsilon_4}\Big)(\tfrac{k}{2}E_1^Q)^2 + \epsilon_4\|v_1\|_0^2 + \|v_0\|_0^2. \end{aligned}$$

Inserting the bounds for E_n^Q from (6.46), we get, using also (A.10),

$$\begin{aligned} k\sum_{n=1}^m E_n^Q \cdot \|\bar{v}_n\|_0 \le \frac{1}{4\epsilon_3}k\sum_{n=2}^m \big(36\|\chi_{n-1}\|_0^2 + 4\|\chi_{n-2}\|_0^2 + Ch^{2(r+q)} \\ + Ck^3\|\ddot{u}\|_{L_2(t_{n-2},t_n;L_2(\Omega))}^2\big) + \epsilon_3 k\sum_{n=2}^m \|\bar{v}_n\|_0^2 \\ + C(1 + \tfrac{1}{\epsilon_4})k^2\big(Ch^{2(r+q)} + 2k^2\|\dot{u}\|_{L_\infty(0,t_1;L_2(\Omega))}^2\big) \\ + \epsilon_4\|v_1\|_0^2 + \|v_0\|_0^2. \end{aligned}$$

Now since $\sum_{n=2}^m \|\ddot{u}\|_{L_2(t_{n-2},t_n;L_2(\Omega))}^2 \le 2\|\ddot{u}\|_{L_2(0,T;L_2(\Omega))}^2$,

$$\begin{aligned} k\sum_{n=1}^m E_n^Q \cdot \|\bar{v}_n\|_0 \le C \cdot (1 + \tfrac{1}{\epsilon_3} + \tfrac{1}{\epsilon_4})h^{2(r+q)} + \tfrac{C}{\epsilon_3}k^4\|\ddot{u}\|_{L_2(0,T;L_2(\Omega))}^2 \\ + C(1 + \tfrac{1}{\epsilon_4})k^4\|\dot{u}\|_{L_\infty(0,t_1;L_2(\Omega))}^2 \\ + \frac{10}{\epsilon_3}k\sum_{n=0}^{m-1}\|\chi_n\|_0^2 + \epsilon_3 k\sum_{n=2}^m \|\bar{v}_n\|_0^2 + \epsilon_4\|v_1\|_0^2 + \|v_0\|_0^2. \end{aligned}$$

To conclude the proof of (6.37) note that by assumption $\|v_0\|_0^2 \le Ch^{2(r+q)}$ and that by (6.11) the summation of $\|\chi_n\|_0^2$ can start at $n = 1$, thus we arrive at (6.37).

To prove the second inequality, (6.38) for $Q = 4$ and $Q = 5$, we also form the term on the left-hand side of (6.38) and use Young's inequality (A.2) to get

$$k\sum_{n=1}^{m} E_n^Q \cdot \|\bar{v}_n\|_0 \leq \frac{1}{4\epsilon}k\sum_{n=1}^{m}(E_n^Q)^2 + \epsilon k\sum_{n=1}^{m}\|\bar{v}_n\|_0^2 \tag{6.47}$$

Now for $Q = 4$ from (6.44), we have $E_n^4 \leq \|\chi_n\|_0 + \|\xi_n\|_0$. Since $\chi_m = 2\bar{\chi}_m - \chi_{m-1}$ we can write

$$\begin{aligned}
\sum_{n=1}^{m}(E_n^4)^2 &\leq 2\sum_{n=1}^{m}\|\chi_n\|_0^2 + 2\sum_{n=1}^{m}\|\xi_n\|_0^2 \\
&\leq 2\|\chi_m\|_0^2 + 2\sum_{n=1}^{m-1}\|\chi_n\|_0^2 + 2\sum_{n=1}^{m}\|\xi_n\|_0^2 \\
&\leq 2(2\|\bar{\chi}_m\|_0 + \|\chi_{m-1}\|_0)^2 + 2\sum_{n=1}^{m-1}\|\chi_n\|_0^2 + 2\sum_{n=1}^{m}\|\xi_n\|_0^2 \\
&\leq 16\|\bar{\chi}_m\|_0^2 + 4\|\chi_{m-1}\|_0^2 + 2\sum_{n=1}^{m-1}\|\chi_n\|_0^2 + 2\sum_{n=1}^{m}\|\xi_n\|_0^2 \\
&\leq 16\|\bar{\chi}_m\|_0^2 + 6\sum_{n=1}^{m-1}\|\chi_n\|_0^2 + 2\sum_{n=1}^{m}\|\xi_n\|_0^2.
\end{aligned}$$

For $Q = 5$ we have from (6.45) $E_n^5 \leq 2(\|\bar{\chi}_n\|_0 + \|\bar{\xi}_n\|_0)$, and we can write, using (5.5)

$$\begin{aligned}
\sum_{n=1}^{m}(E_n^5)^2 &\leq 8\sum_{n=1}^{m}\|\bar{\chi}_n\|_0^2 + 8\sum_{n=1}^{m}\|\bar{\xi}_n\|_0^2 \\
&\leq 8\|\bar{\chi}_m\|_0^2 + 8\sum_{n=1}^{m-1}\|\chi_n\|_0^2 + 4\|\chi_0\|_0^2 + 8\sum_{n=1}^{m}\|\bar{\xi}_n\|_0^2.
\end{aligned}$$

To combine methods $Q = 4$ and $Q = 5$ we observe that for both these methods, after multiplying by k and using (6.11) and (6.3)

$$k\sum_{n=1}^{m}(E_n^Q)^2 \leq 16k\|\bar{\chi}_m\|_0^2 + 8k\sum_{n=1}^{m-1}\|\chi_n\|_0^2 + Ch^{2(r+q)},$$

and inserting this into (6.47) gives

$$k\sum_{n=1}^{m} E_n^Q \cdot \|\bar{v}_n\|_0 \leq \frac{4}{\epsilon}k\|\bar{\chi}_m\|_0^2 + \frac{2}{\epsilon}k\sum_{n=1}^{m-1}\|\chi_n\|_0^2 + \frac{C}{\epsilon}h^{2(r+q)} + \epsilon k\sum_{n=1}^{m}\|\bar{v}_n\|_0^2,$$

which is (6.38).

The regularity assumptions arise because we have used (6.11), (6.12) and (6.3), as well as (5.7) and (5.10) for $Q = 1$ and $Q = 2$. Note that $H^2(0,T;L_2(\Omega))$ is continuously embedded in $W_\infty^1(0,T;L_2(\Omega))$ by (2.22). □

Now the results from the previous lemmas can be combined to get the lower bounds on the term (6.27). Note that in the next lemma the nonlinear methods $Q = 4$ and $Q = 5$ require one extra term in the bound (6.48), as compared to the linearised methods. This term will later give a maximum value for the time step size k for $Q = 4$ and $Q = 5$.

Lemma 6.7 (Nonlinearity errors for methods $Q \in \{1,2,4,5\}$).
Suppose that $\mathcal{ER}^q$ holds, that $\breve{u}, \breve{\sigma} \in H^{r+1}(\Omega)$,

$$u \in L_\infty(0,T;H^{r+1}(\Omega)) \cap W^1_\infty(0,T;L_4(\Omega)) \cap H^2(0,T;L_2(\Omega)) \text{ and}$$
$$\sigma \in L_\infty(0,T;H^{r+1}(\Omega)) \cap W^1_\infty(0,T;L_4(\Omega)).$$

Then for any method $Q \in \{1,2,4,5\}$, there are constants $C, c_3, c_4, c_5^Q, c_6, c_7, c_8, c_9$ such that for any $m \in \{1,\dots,N\}$, we have

$$\begin{aligned} 2k\sum_{n=1}^{m}(\mathscr{B}_n^Q - \overline{\gamma(u)\sigma}_n, \bar{\eta}_n)_0 \geq \check{\gamma}k\sum_{n=1}^{m}\|\bar{\eta}_n\|_0^2 - C(k^4 + h^{2(r+q)}) - c_3\,k\sum_{n=1}^{m-1}\|\chi_n\|_0^2 \\ - c_4\,k\sum_{n=1}^{m-1}\|\eta_n\|_0^2 - c_5^Q\,k\|\chi_m\|_0^2 - \frac{1}{2}\|\eta_1\|_0^2, \end{aligned} \tag{6.48}$$

and also

$$\begin{aligned} 2k\sum_{n=1}^{m}(\mathscr{B}_n^Q - \overline{\gamma(u)\sigma}_n, \bar{\chi}_n + \bar{\eta}_n)_0 \geq -C(k^4 + h^{2(r+q)}) - \frac{1}{2}\|\chi_1 + \eta_1\|_0^2 \\ - c_6\,k\sum_{n=1}^{m-1}\|\chi_n\|_0^2 - c_7\,k\sum_{n=1}^{m-1}\|\eta_n\|_0^2 \\ - c_8\,k\|\bar{\chi}_m\|_0^2 - c_9\,k\|\bar{\eta}_m\|_0^2, \end{aligned} \tag{6.49}$$

where the constants are depending on u and σ, but independent of h and k. In fact $c_1 = C'_\gamma\|\sigma\|_{L_\infty(0,T;L_\infty(\Omega))}$, $c_3 = 28c_1^2/\check{\gamma}$, $c_4 = 8\hat{\gamma}^2/\check{\gamma}$, $c_6 = 4 + 24c_1^2$, $c_7 = 4 + 9\hat{\gamma}^2$, $c_8 = 4 + 8c_1^2$, $c_9 = 4 + \hat{\gamma}^2$ and

$$c_5^Q = \begin{cases} 0 & \text{if } Q \in \{1,2\}, \\ \frac{4(C'_\gamma)^2}{\check{\gamma}}\|\sigma\|^2_{L_\infty(0,T;L_\infty(\Omega))} & \text{if } Q \in \{4,5\}. \end{cases}$$

Proof. To combine the bounds for all methods, we firstly compare (6.28) and (6.32) and thus obtain for any $Q \in \{1,2,4,5\}$ and any $\epsilon_1 > 0$:

$$\begin{aligned} 2k\sum_{n=1}^{m}(\mathscr{B}_n^Q - \overline{\gamma(u)\sigma}_n, \bar{v}_n)_0 \geq 2k\sum_{n=1}^{m}(\gamma(\mathscr{E}_n^Q u^h)\bar{\eta}_n, \bar{v}_n)_0 - \epsilon_1 k\sum_{n=1}^{m}\|\bar{v}_n\|_0^2 \\ - \frac{C}{\epsilon_1}(k^4 + h^{2(r+q)}) - \frac{2c_1^2}{\epsilon_1}k\sum_{n=1}^{m-1}\|\chi_n\|_0^2 - \frac{4\hat{\gamma}^2}{\epsilon_1}k\sum_{n=1}^{m-1}\|\eta_n\|_0^2 \\ - c_1 k\sum_{n=1}^{m}E_n^Q \cdot \|\bar{v}_n\|_0, \end{aligned} \tag{6.50}$$

where E_n^Q is defined as in (6.39). Secondly, by comparing (6.37) and (6.38) we get

$$\begin{aligned} k\sum_{n=1}^{m}E_n^Q \cdot \|\bar{v}_n\|_0 \leq C \cdot (1 + \tfrac{1}{\epsilon_3} + \tfrac{1}{\epsilon_4}) \cdot (k^4 + h^{2(r+q)}) + \epsilon_4\|v_1\|_0^2 \\ + \epsilon_3 k\sum_{n=1}^{m}\|\bar{v}_n\|_0^2 + \frac{10}{\epsilon_3}k\sum_{n=1}^{m-1}\|\chi_n\|_0^2 + \frac{4I^Q}{\epsilon_3}k\|\bar{\chi}_m\|_0^2, \end{aligned}$$

where I^Q is an indicator for the method, which we define as $I^Q = 1$ for $Q \in \{4, 5\}$ and $I^Q = 0$ for $Q \in \{1, 2\}$. Inserting this in (6.50) gives

$$\begin{aligned}2k\sum_{n=1}^{m}(\mathscr{B}_n^Q - \overline{\gamma(u)\sigma}_n, \bar{v}_n)_0 \geq\ & 2k\sum_{n=1}^{m}(\gamma(\mathscr{E}_n^Q u^h)\bar{\eta}_n, \bar{v}_n)_0 - (\epsilon_1 + c_1\epsilon_3)k\sum_{n=1}^{m}\|\bar{v}_n\|_0^2 \\ & - C\cdot(1 + \tfrac{1}{\epsilon_1} + \tfrac{1}{\epsilon_3} + \tfrac{1}{\epsilon_4})(k^4 + h^{2(r+q)}) \\ & - \Big(\frac{2c_1^2}{\epsilon_1} + \frac{10c_1}{\epsilon_3}\Big)k\sum_{n=1}^{m-1}\|\chi_n\|_0^2 - \frac{4\hat{\gamma}^2}{\epsilon_1}k\sum_{n=1}^{m-1}\|\eta_n\|_0^2 \\ & - \frac{4c_1 I^Q}{\epsilon_3}k\|\bar{\chi}_m\|_0^2 - c_1\epsilon_4\|v_1\|_0^2.\end{aligned}$$

Let $\epsilon > 0$ be arbitrary and choose $\epsilon_1 = \epsilon/2$ and $\epsilon_3 = \epsilon/(2c_1)$. Choose also $\epsilon_4 = 1/(2c_1)$. This yields

$$\begin{aligned}2k\sum_{n=1}^{m}(\mathscr{B}_n^Q - \overline{\gamma(u)\sigma}_n, \bar{v}_n)_0 \geq\ & 2k\sum_{n=1}^{m}(\gamma(\mathscr{E}_n^Q u^h)\bar{\eta}_n, \bar{v}_n)_0 - \epsilon k\sum_{n=1}^{m}\|\bar{v}_n\|_0^2 \\ & - C(1 + \tfrac{1}{\epsilon})(k^4 + h^{2(r+q)}) \\ & - \frac{24c_1^2}{\epsilon}k\sum_{n=1}^{m-1}\|\chi_n\|_0^2 - \frac{8\hat{\gamma}^2}{\epsilon}k\sum_{n=1}^{m-1}\|\eta_n\|_0^2 \\ & - \frac{8c_1^2 I^Q}{\epsilon}k\|\bar{\chi}_m\|_0^2 - \frac{1}{2}\|v_1\|_0^2.\end{aligned} \tag{6.51}$$

To prove (6.48) choose $v_n = \eta_n$ in (6.51) and use $(\gamma(\mathscr{E}_n^Q u^h)\bar{\eta}_n, \bar{\eta}_n)_0 \geq \check{\gamma}\|\bar{\eta}_n\|_0^2$. At the same time, to get a term with $\|\chi_m\|_0^2$ instead of $\|\bar{\chi}_m\|_0^2$, we also apply (5.4), i.e. $\|\bar{\chi}_m\|_0^2 \leq \frac{1}{2}(\|\chi_m\|_0^2 + \|\chi_{m-1}\|_0^2)$ and hide the new $\|\chi_{m-1}\|_0^2$ in the sum of $\|\chi_n\|_0^2$:

$$\begin{aligned}2k\sum_{n=1}^{m}(\mathscr{B}_n^Q - \overline{\gamma(u)\sigma}_n, \bar{\eta}_n)_0 \geq\ & (2\check{\gamma} - \epsilon)k\sum_{n=1}^{m}\|\bar{\eta}_n\|_0^2 \\ & - C(1 + \tfrac{1}{\epsilon})(k^4 + h^{2(r+q)}) \\ & - \frac{28c_1^2}{\epsilon}k\sum_{n=1}^{m-1}\|\chi_n\|_0^2 - \frac{8\hat{\gamma}^2}{\epsilon}k\sum_{n=1}^{m-1}\|\eta_n\|_0^2 \\ & - \frac{4c_1^2 I^Q}{\epsilon}k\|\chi_m\|_0^2 - \frac{1}{2}\|\eta_1\|_0^2.\end{aligned}$$

Now choose $\epsilon = \check{\gamma}$ and obtain (6.48).

Next, choose $v_n = \chi_n + \eta_n$ and $\epsilon = 1$ in (6.51) and obtain, since $(\gamma(\mathscr{E}_n^Q u^h)\bar{\eta}_n, \bar{\chi}_n + \bar{\eta}_n)_0 \geq -\hat{\gamma}\|\bar{\eta}_n\|_0 \cdot \|\bar{\chi}_n + \bar{\eta}_n\|_0$,

$$\begin{aligned}2k\sum_{n=1}^{m}(\mathscr{B}_n^Q - \overline{\gamma(u)\sigma}_n, \bar{\chi}_n + \bar{\eta}_n)_0 \geq\ & -2\hat{\gamma}k\sum_{n=1}^{m}\|\bar{\eta}_n\|_0 \cdot \|\bar{\chi}_n + \bar{\eta}_n\|_0 \\ & - k\sum_{n=1}^{m}\|\bar{\chi}_n + \bar{\eta}_n\|_0^2 - C(k^4 + h^{2(r+q)}) \\ & - 24c_1^2 k\sum_{n=1}^{m-1}\|\chi_n\|_0^2 - 8\hat{\gamma}^2 k\sum_{n=1}^{m-1}\|\eta_n\|_0^2 \\ & - 8c_1^2 I^Q k\|\bar{\chi}_m\|_0^2 - \frac{1}{2}\|\chi_1 + \eta_1\|_0^2.\end{aligned} \tag{6.52}$$

Now using Young's inequality (A.1) gives

$$2\hat{\gamma}\|\bar{\eta}_n\|_0 \cdot \|\bar{\chi}_n + \bar{\eta}_n\|_0 \leq \hat{\gamma}^2\|\bar{\eta}_n\|_0^2 + \|\bar{\chi}_n + \bar{\eta}_n\|_0^2,$$

so we can write for the first two terms on the right-hand side of (6.52), using also (5.5),

$$\begin{aligned}
&2\hat{\gamma}k\sum_{n=1}^{m}\|\bar{\eta}_n\|_0 \cdot \|\bar{\chi}_n + \bar{\eta}_n\|_0 + k\sum_{n=1}^{m}\|\bar{\chi}_n + \bar{\eta}_n\|_0^2 \\
&\quad\leq \hat{\gamma}^2 k\sum_{n=1}^{m}\|\bar{\eta}_n\|_0^2 + 2k\sum_{n=1}^{m}\|\bar{\chi}_n + \bar{\eta}_n\|_0^2 \\
&\quad\leq (\hat{\gamma}^2+4)k\sum_{n=1}^{m}\|\bar{\eta}_n\|_0^2 + 4k\sum_{n=1}^{m}\|\bar{\chi}_n\|_0^2 \\
&\quad\leq (\hat{\gamma}^2+4)k\Big(\|\bar{\eta}_m\|_0^2 + \sum_{n=1}^{m-1}\|\eta_n\|_0^2 + \tfrac{1}{2}\|\eta_0\|_0^2\Big) \\
&\qquad + 4k\Big(\|\bar{\chi}_m\|_0^2 + \sum_{n=1}^{m-1}\|\chi_n\|_0^2 + \tfrac{1}{2}\|\chi_0\|_0^2\Big).
\end{aligned}$$

Inserting this into (6.52) and using (6.11) and (6.12) we finally obtain

$$\begin{aligned}
2k\sum_{n=1}^{m}(\mathscr{B}_n^Q - \overline{\gamma(u)\sigma}_n, \bar{\chi}_n + \bar{\eta}_n)_0 \geq\ & -C(k^4 + h^{2(r+q)}) - \frac{1}{2}\|\chi_1 + \eta_1\|_0^2 \\
& - (24c_1^2+4)k\sum_{n=1}^{m-1}\|\chi_n\|_0^2 - (9\hat{\gamma}^2+4)k\sum_{n=1}^{m-1}\|\eta_n\|_0^2 \\
& - (8c_1^2 I^Q + 4)k\|\bar{\chi}_m\|_0^2 - (\hat{\gamma}^2+4)k\|\bar{\eta}_m\|_0^2.
\end{aligned}$$

This is (6.49), the second inequality we wanted to prove. The regularity requirements are obtained from those in Lemmas 6.4, 6.5 and 6.6, we do not distinguish between the methods anymore and just take the maximum requirements. □

6.4 Error estimate

In this section we combine the error inequalities in Lemma 6.3 from Section 6.2 with the nonlinearity errors in Lemma 6.7 from Section 6.3. As it turns out, a combined approach that uses both the inverse Laplacian and the $u+\sigma$-version will remove the condition on the time step size k for the linearised methods.

Theorem 6.8 (An a priori error estimate).
Suppose that $\mathcal{ER}^q$ holds, that $\breve{u}, \breve{\sigma} \in H^{r+1}(\Omega)$ and

$$u, \sigma \in W_\infty^1(0,T;H^{r+1}(\Omega)) \cap H^3(0,T;L_2(\Omega)).$$

For the nonlinear methods $Q=4$ and $Q=5$, we also suppose that k is sufficiently small, i.e.

$$k < \hat{k} = \check{\gamma}\cdot\Big(128c_1^2\max\big(2+4c_1^2+3C_V^2, (4+\hat{\gamma}^2)/\check{\gamma}\big)\Big)^{-1},$$

where $c_1 = C'_\gamma\|\sigma\|_{L_\infty(0,T;L_\infty(\Omega))}$. Then for methods $Q \in \{1,2,4,5\}$, we have the following error estimate

$$\begin{aligned}
&\max_{m\in\{1,\dots,N\}}\big(\|u(t_m) - u_m^h\|_0 + \|\sigma(t_m) - \sigma_m^h\|_0\big) \\
&\qquad + h^q\left(k\sum_{n=1}^{N}\|\bar{u}_n + \bar{\sigma}_n - \bar{u}_n^h - \bar{\sigma}_n^h\|_V^2\right)^{1/2} \leq C(k^2 + h^{r+q}).
\end{aligned} \tag{6.53}$$

Proof. Firstly, use (6.48) in (6.17) which gives, omitting the $\|\mathcal{G}^h\chi_m\|_V$-term,

$$\begin{aligned}
&\|\eta_m\|_0^2 + k\sum_{n=1}^{m}\|\bar{\chi}_n\|_0^2 + \check{\gamma}k\sum_{n=1}^{m}\|\bar{\eta}_n\|_0^2 \\
&\quad\le C(1+\tfrac{1}{\epsilon})(k^4 + h^{2(r+q)}) + \epsilon k\sum_{n=1}^{m}\|\bar{\eta}_n\|_0^2 + c_3 k\sum_{n=1}^{m-1}\|\chi_n\|_0^2 + c_4 k\sum_{n=1}^{m-1}\|\eta_n\|_0^2 \\
&\quad + c_5^Q k\|\chi_m\|_0^2 + \frac{1}{2}\|\eta_1\|_0^2.
\end{aligned}$$

Choosing $\epsilon = \check{\gamma}/2$ gives

$$\begin{aligned}
&\|\eta_m\|_0^2 + k\sum_{n=1}^{m}\|\bar{\chi}_n\|_0^2 + \frac{\check{\gamma}}{2}k\sum_{n=1}^{m}\|\bar{\eta}_n\|_0^2 \\
&\quad\le C(k^4 + h^{2(r+q)}) + c_3 k\sum_{n=1}^{m-1}\|\chi_n\|_0^2 + c_4 k\sum_{n=1}^{m-1}\|\eta_n\|_0^2 + c_5^Q k\|\chi_m\|_0^2 + \frac{1}{2}\|\eta_1\|_0^2.
\end{aligned} \tag{6.54}$$

For $m = 1$, the term $\frac{1}{2}\|\eta_1\|_0^2$ in (6.54) can be absorbed into the left-hand side, and, omitting the other terms on the left-hand side, we obtain

$$\frac{1}{2}\|\eta_1\|_0^2 \le C(k^4 + h^{2(r+q)}) + c_5^Q k\|\chi_1\|_0^2.$$

Inserting this again into (6.54) yields

$$\begin{aligned}
&\|\eta_m\|_0^2 + k\sum_{n=1}^{m}\|\bar{\chi}_n\|_0^2 + \frac{\check{\gamma}}{2}k\sum_{n=1}^{m}\|\bar{\eta}_n\|_0^2 \\
&\quad\le C(k^4 + h^{2(r+q)}) + c_3 k\sum_{n=1}^{m-1}\|\chi_n\|_0^2 + c_4 k\sum_{n=1}^{m-1}\|\eta_n\|_0^2 + c_5^Q k\|\chi_m\|_0^2 + c_5^Q k\|\chi_1\|_0^2.
\end{aligned}$$

Depending on whether $m = 1$ or $m > 1$, this last term can either be combined with the $\|\chi_m\|_0^2$-term or with the sum of $\|\chi_n\|_0^2$. Hence for any $m \ge 1$ we can write:

$$\begin{aligned}
&\|\eta_m\|_0^2 + k\sum_{n=1}^{m}\|\bar{\chi}_n\|_0^2 + \frac{\check{\gamma}}{2}k\sum_{n=1}^{m}\|\bar{\eta}_n\|_0^2 \\
&\quad\le C(k^4 + h^{2(r+q)}) + (c_3 + c_5^Q)k\sum_{n=1}^{m-1}\|\chi_n\|_0^2 + c_4 k\sum_{n=1}^{m-1}\|\eta_n\|_0^2 + 2c_5^Q k\|\chi_m\|_0^2.
\end{aligned} \tag{6.55}$$

To get a second inequality that can be combined with this one, use (6.49) in (6.18), which gives

$$\begin{aligned}
&\|\chi_m + \eta_m\|_0^2 + k\sum_{n=1}^{m}\|\bar{\chi}_n + \bar{\eta}_n\|_V^2 \\
&\qquad\le C(k^4 + h^{2(r+q)}) + c_{10}k\sum_{n=1}^{m-1}\|\chi_n\|_0^2 + c_7 k\sum_{n=1}^{m-1}\|\eta_n\|_0^2 \\
&\qquad\quad + c_{11}k\|\bar{\chi}_m\|_0^2 + c_9 k\|\bar{\eta}_m\|_0^2 + \frac{1}{2}\|\chi_1 + \eta_1\|_0^2.
\end{aligned} \tag{6.56}$$

where the $\|\bar{\chi}_n\|_0$-term from (6.18) has been split up and added to the other terms, also with the use of (6.11), giving $c_{10} = c_6 + 6C_V^2$ and $c_{11} = c_8 + 6C_V^2$.

For $m = 1$, the term $\frac{1}{2}\|\chi_1+\eta_1\|_0^2$ in (6.56) cancels with the left-hand side, and, omitting the other term on the left-hand side, we obtain

$$\frac{1}{2}\|\chi_1 + \eta_1\|_0^2 \le C(k^4 + h^{2(r+q)}) + c_{11}k\|\bar{\chi}_1\|_0^2 + c_9 k\|\bar{\eta}_1\|_0^2.$$

Inserting this again into (6.56) yields

$$\begin{aligned}\|\chi_m + \eta_m\|_0^2 + k\sum_{n=1}^{m}\|\bar{\chi}_n + \bar{\eta}_n\|_V^2 \\ \le C(k^4 + h^{2(r+q)}) + c_{10}k\sum_{n=1}^{m-1}\|\chi_n\|_0^2 + c_7 k\sum_{n=1}^{m-1}\|\eta_n\|_0^2 \\ + c_{11}k\|\bar{\chi}_m\|_0^2 + c_9 k\|\bar{\eta}_m\|_0^2 + c_{11}k\|\bar{\chi}_1\|_0^2 + c_9 k\|\bar{\eta}_1\|_0^2.\end{aligned}$$

For convenience of having to write fewer terms we can combine some of them again, enlarging the right-hand side,

$$\begin{aligned}\|\chi_m + \eta_m\|_0^2 + k\sum_{n=1}^{m}\|\bar{\chi}_n + \bar{\eta}_n\|_V^2 \\ \le C(k^4 + h^{2(r+q)}) + c_{10}k\sum_{n=1}^{m-1}\|\chi_n\|_0^2 + c_7 k\sum_{n=1}^{m-1}\|\eta_n\|_0^2 \\ + 2c_{11}k\sum_{n=1}^{m}\|\bar{\chi}_n\|_0^2 + 2c_9 k\sum_{n=1}^{m}\|\bar{\eta}_n\|_0^2.\end{aligned} \tag{6.57}$$

Now we define $c_{12} = \max(2c_{11}, 4c_9/\check{\gamma})$. Multiplying (6.55) by c_{12} and adding this to (6.57) gives,

$$\begin{aligned}\|\chi_m + \eta_m\|_0^2 + c_{12}\|\eta_m\|_0^2 + c_{12}k\sum_{n=1}^{m}\|\bar{\chi}_n\|_0^2 + c_{12}\frac{\check{\gamma}}{2}k\sum_{n=1}^{m}\|\bar{\eta}_n\|_0^2 + k\sum_{n=1}^{m}\|\bar{\chi}_n + \bar{\eta}_n\|_V^2 \\ \le C(k^4 + h^{2(r+q)}) + (c_{10} + c_{12}(c_3 + c_5^Q))k\sum_{n=1}^{m-1}\|\chi_n\|_0^2 + (c_7 + c_{12}c_4)k\sum_{n=1}^{m-1}\|\eta_n\|_0^2 \\ + 2c_{11}k\sum_{n=1}^{m}\|\bar{\chi}_n\|_0^2 + 2c_9 k\sum_{n=1}^{m}\|\bar{\eta}_n\|_0^2 + 2c_{12}c_5^Q k\|\chi_m\|_0^2.\end{aligned}$$

Since $2c_{11} \le c_{12}$ and $2c_9 \le c_{12}\check{\gamma}/2$, we can eliminate more terms on the right-hand side:

$$\begin{aligned}\|\chi_m + \eta_m\|_0^2 + c_{12}\|\eta_m\|_0^2 + k\sum_{n=1}^{m}\|\bar{\chi}_n + \bar{\eta}_n\|_V^2 \\ \le C(k^4 + h^{2(r+q)}) + Ck\sum_{n=1}^{m-1}(\|\chi_n\|_0^2 + \|\eta_n\|_0^2) + 2c_{12}c_5^Q k\|\chi_m\|_0^2.\end{aligned} \tag{6.58}$$

If we set $c_{13} = \max(4, 5/c_{12})$, then

$$\begin{aligned}2\|\chi_m\|_0^2 + \|\eta_m\|_0^2 \le 2(\|\chi_m + \eta_m\|_0 + \|\eta_m\|_0)^2 + \|\eta_m\|_0^2 \le 4\|\chi_m + \eta_m\|_0^2 + 5\|\eta_m\|_0^2 \\ \le 4\|\chi_m + \eta_m\|_0^2 + \frac{5}{c_{12}}c_{12}\|\eta_m\|_0^2 \le c_{13}(\|\chi_m + \eta_m\|_0^2 + c_{12}\|\eta_m\|_0^2),\end{aligned}$$

and hence $\|\chi_m + \eta_m\|_0^2 + c_{12}\|\eta_m\|_0^2 \geq c_{13}^{-1}(2\|\chi_m\|_0^2 + \|\eta_m\|_0^2)$. Using this in (6.58) and multiplying by c_{13} gives

$$\begin{aligned}
&2\|\chi_m\|_0^2 + \|\eta_m\|_0^2 + c_{13}k\sum_{n=1}^{m}\|\bar{\chi}_n + \bar{\eta}_n\|_V^2 \\
&\qquad \leq C(k^4 + h^{2(r+q)}) + Ck\sum_{n=1}^{m-1}(\|\chi_n\|_0^2 + \|\eta_n\|_0^2) + 2c_{13}c_{12}c_5^Q k\|\chi_m\|_0^2.
\end{aligned}$$

For the methods $Q = 4$ and $Q = 5$, where $c_5^Q > 0$, we now need that $k \leq \hat{k} = 1/(2c_{13}c_{12}c_5^Q)$, so that the term $\|\chi_m\|_0^2$ on the right-hand side can be absorbed into the left-hand side. Hence, for all $Q = 1, 2, 4, 5$ we get

$$\begin{aligned}
&\|\chi_m\|_0^2 + \|\eta_m\|_0^2 + c_{13}k\sum_{n=1}^{m}\|\bar{\chi}_n + \bar{\eta}_n\|_V^2 \\
&\qquad \leq C(k^4 + h^{2(r+q)}) + Ck\sum_{n=1}^{m-1}(\|\chi_n\|_0^2 + \|\eta_n\|_0^2).
\end{aligned}$$

Using the discrete Grönwall lemma, and multiplying by a suitable constant, we get

$$\|\chi_m\|_0^2 + \|\eta_m\|_0^2 + k\sum_{n=1}^{m}\|\bar{\chi}_n + \bar{\eta}_n\|_V^2 \leq C(k^4 + h^{2(r+q)}). \tag{6.59}$$

Observe that the term with the V-norm does not need a factor of h^{2q} here. But for the final step of the proof, we have to introduce this factor to get the bound for ξ_n and θ_n that corresponds to (6.59): From Lemma 6.2 we obtain

$$\|\xi_m\|_0^2 + \|\theta_m\|_0^2 + h^{2q}\,k\sum_{n=1}^{m}\|\bar{\xi}_n + \bar{\theta}_n\|_V^2 \leq Ch^{2(r+q)}. \tag{6.60}$$

In order to introduce the factor h^{2q} into (6.59), we note that we have an upper bound for h, $h \leq \hat{h}$ and using (6.59) gives us

$$\begin{aligned}
&\|\chi_m\|_0^2 + \|\eta_m\|_0^2 + h^{2q}k\sum_{n=1}^{m}\|\bar{\chi}_n + \bar{\eta}_n\|_V^2 \\
&\qquad \leq \max(1, \hat{h}^{2q})\Big(\|\chi_m\|_0^2 + \|\eta_m\|_0^2 + k\sum_{n=1}^{m}\|\bar{\chi}_n + \bar{\eta}_n\|_V^2\Big) \\
&\qquad \leq C(k^4 + h^{2(r+q)}).
\end{aligned} \tag{6.61}$$

Finally since $u_n - u_n^h = \xi_n - \chi_n$ and $\sigma_n - \sigma_n^h = \theta_n - \eta_n$ we get by using the triangle inequality, (6.61) and (6.60)

$$\begin{aligned}
&\|u_m - u_m^h\|_0^2 + \|\sigma_m - \sigma_m^h\|_0^2 + h^{2q}\,k\sum_{n=1}^{m}\|\bar{u}_n + \bar{\sigma}_n - \bar{u}_n^h - \bar{\sigma}_n^h\|_V^2 \\
&\leq 2\|\xi_m\|_0^2 + 2\|\theta_m\|_0^2 + 2h^{2q}\,k\sum_{n=1}^{m}\|\bar{\xi}_n + \bar{\theta}_n\|_V^2 \\
&\quad + 2\|\chi_m\|_0^2 + 2\|\eta_m\|_0^2 + 2h^{2q}\,k\sum_{n=1}^{m}\|\bar{\chi}_n + \bar{\eta}_n\|_V^2 \\
&\leq C(k^4 + h^{2(r+q)}),
\end{aligned}$$

where we have used (6.59) and again Lemma 6.2. To get the maximum time step for $Q \in \{4, 5\}$ we note that

$$\begin{aligned}\frac{1}{\hat{k}} &= 2c_5^Q c_{13} c_{12} = 2c_5^Q \max(4c_{12}, 5) = 2c_5^Q \max(5, 8c_{11}, 16c_9/\check{\gamma}) \\ &= \frac{8c_1^2}{\check{\gamma}} \max(5, 8c_{11}, 16c_9/\check{\gamma}) = \frac{8c_1^2}{\check{\gamma}} \max\Big(5, 32 + 64c_1^2 + 48C_V^2, \frac{64 + 16\hat{\gamma}^2}{\check{\gamma}}\Big).\end{aligned}$$

Rearranging gives the expression for $\hat{k}$. □

6.4.1 Why so complicated?

Here, we want to give a small explanation why the proof of the error estimate uses both the inverse Laplacian *and* the '$u + \sigma$'-approach.

Inverse Laplacian

Looking back at the proof of Theorem 6.8 we see that when using only the inverse Laplacian we end up with (6.55). If we do not omit the term $\|\mathcal{G}^h \chi_m\|_V^2$, as we did in the beginning of the proof, and if we restrict ourselves to the linearised methods $Q = 1$ and 2, then (6.55) looks like this:

$$\begin{aligned}&\|\mathcal{G}^h \chi_m\|_V^2 + \|\eta_m\|_0^2 + k\sum_{n=1}^{m} \|\bar{\chi}_n\|_0^2 + \frac{\check{\gamma}}{2} k \sum_{n=1}^{m} \|\bar{\eta}_n\|_0^2 \\ &\qquad \le C(k^4 + h^{2(r+1)}) + Ck \sum_{n=1}^{m-1} \|\chi_n\|_0^2 + Ck \sum_{n=1}^{m-1} \|\eta_n\|_0^2.\end{aligned} \tag{6.62}$$

Using Grönwall's lemma, we can eliminate the term with $\|\eta_n\|_0$ from the right-hand side. However, since $\|\mathcal{G}^h \chi_m\|_V \le \|\chi_m\|_{V'}$, the term $\|\mathcal{G}^h \chi_m\|_V$ is not strong enough to eliminate the term with $\|\chi_n\|_0$ from the right-hand side using Grönwall's lemma. Thus, we need $\|\chi_m\|_0^2$ instead of $\|\mathcal{G}^h \chi_m\|_V^2$ on the left-hand side, but the inverse Laplacian does not provide this term.

'$u + \sigma$'-approach

On the other hand, using the '$u+\sigma$'-approach, we would have (6.57), to which we would add an inequality that comes from the σ-equation only: From (6.22b), we obtain with the usual manipulations and by using (6.49), again restricting ourselves to $Q \in \{1, 2\}$,

$$\|\eta_m\|_0^2 + k\sum_{n=1}^{m} \|\bar{\eta}_n\|_0^2 \le C(k^4 + h^{2(r+1)}) + Ck \sum_{n=1}^{m} \|\bar{\chi}_n\|_0^2.$$

Combining this with (6.57) and using (4.7) we arrive at

$$\begin{aligned}&\|\chi_m\|_0^2 + \|\eta_m\|_0^2 + k\sum_{n=1}^{m} \|\bar{\chi}_n + \bar{\eta}_n\|_V^2 + k\sum_{n=1}^{m} \|\bar{\eta}_n\|_0^2 \\ &\qquad \le C(k^4 + h^{2(r+1)}) + Ck\|\chi_m\|_0^2 + Ck\|\eta_m\|_0^2 + Ck \sum_{n=1}^{m-1} \|\chi_n\|_0^2 + Ck \sum_{n=1}^{m-1} \|\eta_n\|_0^2.\end{aligned} \tag{6.63}$$

Here, the sums of $\|\chi_n\|_0^2$ and $\|\eta_n\|_0^2$ can easily be eliminated using Grönwall's lemma. However in order to eliminate $\|\chi_m\|_0^2$ and $\|\eta_m\|_0^2$ we would need that k is small enough to be able to cancel it against the first terms on the left-hand side.

Combination of both

The idea of the proof of Theorem 6.8 was to combine both approaches by adding (6.62), multiplied by a big enough factor, to (6.63), so that the terms $Ck\|\chi_m\|_0^2$ and $Ck\|\eta_m\|_0^2$ from the right-hand side can be cancelled against the terms $k\sum_{n=1}^m\|\bar\chi_n\|_0^2$ and $k\sum_{n=1}^m\|\bar\eta_n\|_0^2$ on the left-hand side of (6.62). (For this, we have to bound $\|\chi_m\|_0 \le 2\|\bar\chi_m\|_0 + \|\chi_{m-1}\|_0$ and include $\|\chi_{m-1}\|_0$ in the sum of $\|\chi_n\|_0$.) Now, the norm of χ_m on the left-hand side is also strong enough for Grönwall's lemma and the remaining terms on the right-hand side can be eliminated.

6.5 About method $Q = 3$

This section contains some thoughts on how an error estimate like Theorem 6.8 could be proved for method $Q = 3$. What is needed is a version of the nonlinearity errors in Lemma 6.7 for $Q = 3$.

A possible approach is to use the results for $Q = 1$, since $\mathscr{B}_n^3 = I^h\mathscr{B}_n^1$. The error of the nonlinearity can be split into two parts:

$$\mathscr{B}_n^3 - \overline{\gamma(u)\sigma}_n = (I^h\mathscr{B}_n^1 - \mathscr{B}_n^1) + (\mathscr{B}_n^1 - \overline{\gamma(u)\sigma}_n). \tag{6.64}$$

For the second part, $\mathscr{B}_n^1 - \overline{\gamma(u)\sigma}_n$, the bounds from Lemma 6.7 can be used. To control the first part, $I^h\mathscr{B}_n^1 - \mathscr{B}_n^1$, an estimate like

$$k\sum_{n=1}^m\|I^h\mathscr{B}_n^1 - \mathscr{B}_n^1\|_0^2 \le Ch^{2(r+q)}. \tag{6.65}$$

would be sufficient to prove Lemma 6.7 for $Q = 3$, the only difference is that the constants in the inequalities (6.48) and (6.49) are changed:

Lemma 6.9 (Tentative nonlinearity error for $Q = 3$).
Suppose that (6.65) *holds. Then, with the assumptions and definitions of constants from Lemma 6.7, we have for any $m \in \{1, \dots, N\}$,*

$$\begin{aligned} 2k\sum_{n=1}^m(\mathscr{B}_n^3 - \overline{\gamma(u)\sigma}_n, \bar\eta_n)_0 \ge{}& \frac{\check\gamma}{2}k\sum_{n=1}^m\|\bar\eta_n\|_0^2 - C(k^4 + h^{2(r+q)}) - c_3\,k\sum_{n=1}^{m-1}\|\chi_n\|_0^2 \\ &- c_4\,k\sum_{n=1}^{m-1}\|\eta_n\|_0^2 - \frac{1}{2}\|\eta_1\|_0^2, \end{aligned} \tag{6.66}$$

and also

$$\begin{aligned} 2k\sum_{n=1}^m(\mathscr{B}_n^3 - \overline{\gamma(u)\sigma}_n, \bar\chi_n + \bar\eta_n)_0 \ge{}& -C(k^4 + h^{2(r+q)}) - \frac{1}{2}\|\chi_1 + \eta_1\|_0^2 \\ &- (2 + c_6)\,k\sum_{n=1}^{m-1}\|\chi_n\|_0^2 - (2 + c_7)\,k\sum_{n=1}^{m-1}\|\eta_n\|_0^2 \\ &- (2 + c_8)\,k\|\bar\chi_m\|_0^2 - (2 + c_9)\,k\|\bar\eta_m\|_0^2. \end{aligned} \tag{6.67}$$

Proof. Using (6.64) to split the error, gives

$$\begin{aligned}2k\sum_{n=1}^{m}(\mathscr{B}_n^3-\overline{\gamma(u)\sigma}_n,\bar{v}_n)_0 &\geq 2k\sum_{n=1}^{m}(\mathscr{B}_n^1-\overline{\gamma(u)\sigma}_n,\bar{v}_n)_0-2k\sum_{n=1}^{m}\|I^h\mathscr{B}_n^1-\mathscr{B}_n^1\|_0\|\bar{v}_n\|_0\\ &\geq 2k\sum_{n=1}^{m}(\mathscr{B}_n^1-\overline{\gamma(u)\sigma}_n,\bar{v}_n)_0\\ &\qquad-\frac{1}{\epsilon}k\sum_{n=1}^{m}\|I^h\mathscr{B}_n^1-\mathscr{B}_n^1\|_0^2-\epsilon k\sum_{n=1}^{m}\|\bar{v}_n\|_0^2 \qquad (6.68)\end{aligned}$$

Now use Lemma 6.7 for $Q=1$ and the assumption (6.65) to finish the proof: Firstly choose $\bar{v}_n=\bar{\eta}_n$ and $\epsilon=\check{\gamma}/2$ in (6.68) to obtain (6.66). Secondly choose $\bar{v}_n=\bar{\chi}_n+\bar{\eta}_n$ and $\epsilon=1$ in (6.68), then split the sum

$$\begin{aligned}&k\sum_{n=1}^{m}\|\bar{\chi}_n+\bar{\eta}_n\|_0^2\\ &\qquad\leq 2k\|\bar{\chi}_m\|_0^2+2k\|\bar{\eta}_m\|_0^2+2k\sum_{n=1}^{m-1}\|\chi_n\|_0^2+2k\sum_{n=1}^{m-1}\|\eta_n\|_0^2+2k\|\chi_0\|_0^2+2k\|\eta_0\|_0^2,\end{aligned}$$

and add the resulting terms to those which come from (6.49), using also (6.11) and (6.12). This yields (6.67). □

The changed sizes of the constants do not matter for the proof of the error estimate. Thus, if (6.65) holds under certain conditions, the error estimate will follow from that. But proving (6.65) seems to be difficult. The obvious way to start is to use the approximation error of the interpolation operator I^h, Lemma 2.29:

$$\|I^h\mathscr{B}_n^1-\mathscr{B}_n^1\|_0\leq Ch^{r+q}\|\mathscr{B}_n^1\|_{H^{r+q}(\Omega)},$$

if $\mathscr{B}_n^1\in H^{r+q}(\Omega)$ and $r+q>d/2$. This immediately implies

$$k\sum_{n=1}^{m}\|I^h\mathscr{B}_n^1-\mathscr{B}_n^1\|_0^2\leq Ch^{2(r+q)}k\sum_{n=1}^{m}\|\mathscr{B}_n^1\|_{H^{r+q}(\Omega)}^2. \qquad (6.69)$$

Hence, provided that we can bound $\mathscr{B}_n^1\in H^{r+q}(\Omega)$ in terms of the problem data, we would get the required bound (6.65).

Now even in the simplest case of $d=1$, $r=1$ and $q=0$ we need a bound for

$$\|\mathscr{B}_n^1\|_{H^1(\Omega)},$$

depending only on the problem data, or equivalently a bound for $\|\mathscr{B}_n^1\|_V$. The only bound that we have and that comes close to this, however is the bound for

$$\|u_n^h+\sigma_n^h\|_V$$

from Proposition 5.18. This appears not to be enough for obtaining (6.65) and the problem of proving an error estimate for $Q=3$ remains open.

Chapter 7

Numerical Results

In this chapter we present numerical results for the discretisations that were given in the previous chapters. We want to show that we can get the convergence rates that are predicted from the theory.

7.1 Another discrete formulation

We recall that in Chapter 5 we introduced the discrete formulation of Problem 5.4 (page 62), which was based on the simplified variational formulation of Problem 4.3 (page 43) and was used as the basis for the theoretical results.

In this section we present another discrete formulation. It is based on the variational formulation of Problem 4.1 (page 42). Both discrete formulations were implemented and we present numerical results for both of them.

Problem 7.1 (Alternative discrete formulation).
For $n = 1, \dots, N$ find discrete solutions $u_n^h \in V^h + I^h u_D(t_n) \subset W^h$ and $\sigma_n^h \in W^h$ such that

$$(\partial_t u_n^h, v)_0 + D(\bar{u}_n^h, v)_V + K(\bar{\sigma}_n^h, v)_V = \langle \bar{L}_n, v \rangle \qquad \forall\, v \in V^h,\ n = 1, \dots, N, \tag{7.1a}$$

$$(\partial_t \sigma_n^h, w)_0 + (\mathscr{B}_n^Q, w)_0 = \mu(\bar{u}_n^h, w)_0 + (\bar{\hat{f}}_n, w)_0 \qquad \forall\, w \in W^h,\ n = 1, \dots, N, \tag{7.1b}$$

$$(u_0^h, w)_0 = (\breve{u}, w)_0 \qquad \forall\, w \in W^h, \tag{7.1c}$$

$$(\sigma_0^h, w)_0 = (\breve{\sigma}, w)_0 \qquad \forall\, w \in W^h. \tag{7.1d}$$

where $\mathscr{B}_n^Q$ is defined as in Definition 5.5.

The differences to Problem 5.4 are that here, no Dirichlet boundary conditions are imposed on σ, the initial discrete solutions are the L_2-projection into W^h and not into V^h and that the Dirichlet boundary conditions on u can be inhomogeneous.

In fact, Problem 5.4 was also implemented for inhomogeneous Dirichlet boundary conditions on u.

For having short notation in the rest of this chapter, we define the parameter $\sigma_D \in \{0, 1\}$ which indicates whether Dirichlet boundary conditions for σ are prescribed:

$$\sigma_D := \begin{cases} 0 & \text{if Problem 7.1 is solved,} \\ 1 & \text{if Problem 5.4 is solved.} \end{cases} \tag{7.2}$$

7.2 Implementation

The computer program that was written for the implementation uses the finite element library ALBERTA 1.2, see [SS05]. The implementation covers:

- both the simplified discrete Problem 5.4 (which prescribes Dirichlet boundary conditions for σ) and Problem 7.1,
- in all cases $\nu = 0$ and the parameters D, K, μ are constants, not dependent on u,
- all three versions of γ from Subsection 3.5.1,
- all five methods of discretisation for the nonlinearity $Q \in \{1, 2, 3, 4, 5\}$,
- all space dimensions $d \in \{1, 2, 3\}$,
- all polynomial degrees of Lagrange basis functions that are provided by ALBERTA: $r \in \{1, 2, 3, 4\}$.

The ALBERTA library also provides functionality for time and space adaptivity. This means that the size of the time step can change and that the size of the finite elements is based on local estimates of the error. But this functionality is not used in this thesis. The program uses a fixed time step and the mesh is also fixed for the whole of each calculation.

For $Q \in \{1, 2, 4, 5\}$ the discrete solution is found using the matrix formulation from Lemma 5.7 (and a similar matrix formulation for Problem 7.1). As the linear system is non-symmetric, GMRes was chosen as a linear solver, see [SS05].

For $Q = 3$ the smaller linear system from Lemma 5.8 is solved. Here the conjugate gradient method (CG) was chosen as linear solver. This solver has the advantage of being much faster than GMRes to achieve the desired accuracy. The disadvantage is that it can normally only be applied to symmetric matrices. The system matrix for $Q = 3$ is still in general not symmetric but for practical purposes it seems to be close enough to being symmetric. To make sure that the solutions which are calculated by the CG method are accurate enough, the residual was checked.

For the nonlinear methods $Q \in \{4, 5\}$, a nonlinear system has to be solved at each time step. This was implemented using a simple fixed point iteration, similar to the iteration set up for proving existence of an exact solution in Lemma 4.12: The old iterate for u_n^h is used to calculate $\gamma(u_n^h)$ (for $Q = 4$) or $\gamma(\bar{u}_n^h)$ (for $Q = 5$). This gives a linear system which is solved to obtain the new iterate for u_n^h and σ_n^h. The stopping criteria for the fixed point iteration are that either the number of iterations exceeds 1000 or that the residual of the nonlinear system becomes less than 10^{-13}. For this reason the convergence tables also give the values of the residual.

Version 1.2 of the ALBERTA library uses GRAPE ([SFB97]) for visualisation. The plots of two- and three-dimensional numerical solutions of this chapter were created using GRAPE. For the one-dimensional numerical solutions, however, and for other 2D graphs, Gnuplot was used.

7.2.1 Defining errors

For verifying the convergence results of the a priori error estimate (Theorem 6.8) define the following two measures for the error:

$$e_0 := \max_{m\in\{1,\dots,N\}} \big(\|u(t_m) - u_m^h\|_0 + \|\sigma(t_m) - \sigma_m^h\|_0\big), \tag{7.3}$$

$$e_1 := \left(k\sum_{n=1}^{N}\|\bar{u}_n + \bar{\sigma}_n - \bar{u}_n^h - \bar{\sigma}_n^h\|_V^2\right)^{1/2}. \tag{7.4}$$

Using these two definitions we also define the combined error

$$e_{\text{comb}} := e_0 + h\cdot e_1. \tag{7.5}$$

Note that we define the combined error as if the elliptic regularity $\mathcal{ER}^1$ holds in Theorem 6.8. This is done irrespective of whether the assumptions of $\mathcal{ER}^1$ are satisfied for the numerical experiments of this chapter.

7.3 Testing the programs for exactness

To have an indication that the computer implementation works correctly, the *exactness* of the program was tested, i.e. it was checked whether the numerical solution coincides with the exact solution whenever it should.

From the way the discrete solution is constructed, by comparing Problem 4.3 with Problem 5.4 we see that the numerical solution should coincide with the exact solution, if

$$\begin{aligned} u_n, \sigma_n &\in V^h \text{ (respectively } W^h \text{ for } \sigma_D = 0),\\ \bar{\dot{u}} &= (u_n - u_{n-1})/k,\\ \bar{\dot{\sigma}} &= (\sigma_n - \sigma_{n-1})/k,\\ \overline{\gamma(u)\sigma}_n &= \mathscr{B}_n^Q(u,\sigma). \end{aligned}$$

The conditions on the time derivatives are satisfied if u and σ are globally polynomials in t of degree at most 2. The last condition depends on Q: For $Q = 4$, we get equality in any case. If u is constant in time, we get equality for $Q = 1, 2, 5$. If u is constant in time and space, we get equality for $Q = 3$.

For creating test problems, choose functions u and σ and calculate

$$\begin{aligned} f &= \dot{u} - \Delta u - \Delta\sigma,\\ \hat{f} &= \dot{\sigma} + \gamma(u)\sigma - u,\\ g &= (\nabla u + \nabla\sigma)\cdot \boldsymbol{n}_{\partial\Omega}. \end{aligned}$$

This way the discrete solution can be compared with the exact solution.

Choosing test problems with these properties, the error turns out to be almost vanishing except in two cases:

- quartic polynomials in 3D: $r = 4$, $d = 3$, see Subsection 7.3.1,
- $Q \in \{4, 5\}$, see later in Subsection 7.4.2.

7.3.1 Degree 4 polynomials in 3D

In the case where $r = 4$ and $d = 3$, the exactness appears to be lost. This seems to be due to the mass matrix, which for one element is a 35 by 35 matrix that appears not to be practically invertible anymore. The condition number of this matrix with respect to the 1-norm has an approximate order of magnitude of 10^{10}.

7.4 Convergence tables

To demonstrate the error estimate of Theorem 6.8 in practice, this section contains tables of convergence.

For the domain, take the d-dimensional 'cube' $\Omega = (0,1)^d$, with Dirichlet boundary at $x_1 = 0$, i.e. $\Gamma_D = \{0\} \times [0,1]^{d-1} = \{\boldsymbol{x} \in \mathbb{R}^d \ : \ x_1 = 0 \wedge x_2, \ldots, x_d \in [0,1]\}$, and the remainder of the boundary with Neumann conditions, $\Gamma_N = \partial\Omega \setminus \Gamma_D$. Thus we have both a Dirichlet and a Neumann part of the boundary. The time interval is $[0,T] = [0,1]$. Unless stated otherwise the following values were taken for the numerical results of this section: $\check{\gamma} = 1, \hat{\gamma} = 100, u_a = 0.5$. To make the shape of γ_3 and γ_1 comparable, different values of Δ_a were chosen: $\Delta_a = 0.2$ for γ_3 and $\Delta_a = 0.05$ for γ_1. Taking the same value of Δ_a would lead to very different sizes of the (approximate) transition region, as can be seen from comparing Figures 3.1 and 3.3.

Convergence for time discretisation

The problem considered for looking at the time discretisation for $Q \in \{1,2,4,5\}$ has u and σ nonlinear in time and linear in space:

$$u = \cos(t) \sum_{i=1}^{d} x_i,$$
$$\sigma = t^3 \sum_{i=1}^{d} (d - i + 1)\, x_i.$$

For method $Q = 3$, we take a simpler test problem:

$$u = \cos(t),$$
$$\sigma = t^3 \sum_{i=1}^{d} (d - i + 1)\, x_i.$$

These problems are chosen so that the error of space discretisation vanishes and the k-dependence of the error can be demonstrated.

Convergence for spatial discretisation

The problem considered for looking at the space discretisation has u and σ nonlinear in space and σ linear in time:

$$u = \prod_{l=1}^{d} \sin(\pi x_l),$$
$$\sigma = (1 - t)\, \exp\left(\sum_{l=1}^{d} x_l \right).$$

For this problem, the error of time discretisation vanishes and the h-dependence of the error can be seen.

Explanation of table headings

The convergence tables all have a similar structure, depending on whether they show h- or k-convergence. The following variables are tabulated:

- N: the number of time steps
- k: the time step size
- M: a number representing the fineness of the mesh such that $h = C/M$. The mesh consists of M^d "building blocks" of elements, by subdividing the domain along each coordinate axis into M parts. In 1D each building block is an element. In 2D each (square) building block is divided into four triangular elements, as in ⊠. In 3D each (cube-shaped) building block is divided into six tetrahedral elements.
- h: maximum element diameter
- ndof: the number of degrees of freedom, i.e. the value N_W. For $Q = 3$ this is the size of the linear system, whereas for the other values of Q this is half the size of the linear system.
- residual: This is the maximum (over time) of the Euclidean norms of the residuals that occurred solving the linear or nonlinear systems.
- Errors e_0, e_1 and e_{comb} as defined in (7.3), (7.4) and (7.5). As these three errors do not fit all into one table, two tables are given for each combination of parameters: one with e_0 and e_1 and another one with e_{comb}.
- EOC: the *estimated order of convergence* for the error in the previous column. If the previous row had an error of e_{previous} for h_{previous} and the current row has an error of e for h, then
$$\text{EOC} = \frac{\log(e/e_{\text{previous}})}{\log(h/h_{\text{previous}})}. \tag{7.6}$$
For the k-convergence tables, replace h by k in this definition.

7.4.1 Overview of tables

There is a vast choice of possible combinations of problem and discretisation parameters (two choices for σ_D, three choices for γ, five for Q, three for d, four for r giving 360 possibilities). Moreover, even after rescaling, the open parameters $\check{\gamma}$, $\hat{\gamma}$, u_a, Δ_a, the size of the domain Ω, the size of the time interval $[0, T]$ and of course the choice of exact solution can have an effect on the errors.

As it seems impossible to give convergence evidence for all possible combinations, we consider the following choices to be reasonable: The tables concentrate on the case $d = 2$ and $r = 1$. For these, all five methods Q are examined. An important factor could be whether Dirichlet boundary values are prescribed or not, therefore both cases for σ_D are investigated. Another question is whether taking γ_3 instead of the infinitely differentiable γ_1 has an influence.

All other choices of parameters are narrowly selected to keep the number of tables manageable.

We now give an overview of the tables:

- Linear polynomials in 2D: $d = 2$, $r = 1$. All combinations of $\sigma_D \in \{1, 0\}$, $\gamma \in \{\gamma_3, \gamma_1\}$ and $Q \in \{1, 2, 3, 4, 5\}$ are considered. For the k-convergence, see Tables 7.1 to 7.40. For h-convergence, see Tables 7.45 to 7.84.

- 1D and 3D for $\sigma_D = 1$, $\gamma = \gamma_3$, $Q = 1$, $r = 1$. For k-convergence, see Tables 7.41 to 7.44. For h-convergence, see Tables 7.85 to 7.88.
- Higher degree polynomials in 1D: $r \in \{2, 3, 4\}$ for $\sigma_D = 1$, $\gamma = \gamma_3$, $Q = 3$. For h-convergence, see Tables 7.89 to 7.94.
- Examples for the nonlinear methods $Q \in \{4, 5\}$, with tables for $k = 0.0001$ and again for $k = 0.25$ which is too big, see Tables 7.95 to 7.102. Here it can be seen that k needs to be small enough to have h-convergence.
- An example for k-convergence with $\hat{\gamma} = 10$ instead of 100, see Tables 7.103 and 7.104.

To find a specific table, we also refer the reader to the List of Tables at the beginning of the thesis.

For the k-convergence tables, we choose $M = 1$.

7.4.2 Influence of time step k on exactness and h-convergence for $Q \in \{4, 5\}$

For the nonlinear methods exactness and convergence requires that k is small enough. Compare Section 5.4, where we saw that the discrete solution may not be unique for these methods, if the time step is too big.

To demonstrate this for exactness, we consider a problem where $\Omega = (0, 1)$, $\Gamma_D = \{0\}$ and the exact solution is given by

$$u(x, t) = tx, \quad \sigma(x, t) = 2.3.$$

The parameters of γ are $\hat{\gamma} = 2$, $\check{\gamma} = 1$, $u_a = 1, \Delta_a = 0.2$. Solving this with $Q = 4$ should show exactness, but as it turns out, this is only exact for small enough time steps. See Figure 7.1 for the numerical solution using $N = 1, 2, 4, 8$ and 16 time steps.

To demonstrate the effect on the h-convergence, we look at four convergence tables, two of them for $Q = 4$ and another two for $Q = 5$. It can be seen, that the time step needs to be small enough to observe convergence.

$Q = 4$:
For $k = 0.25$ see Table 7.97, while for $k = 0.0001$ see Table 7.95.

$Q = 5$:
For $k = 0.25$ see Table 7.101, while for $k = 0.0001$ see Table 7.99.

As some matrix entries have the factor $1/k$, the system becomes difficult to solve exactly for $k = 0.0001$, but the residuals are still small enough to see the convergence.

7.4.3 Summary of the tables

k-convergence

In the tables for k-convergence (Tables 7.1-7.44) the EOC for all three errors e_0, e_1 and e_{comb} is close to 2. Note however that in some of these cases the EOC of 2 appears only in the last few lines of the table. In the lines before, that is for bigger values of k, the EOC is smaller. This is the case in Tables 7.1-7.2, for example. This effect seems to be due to the choice of $\hat{\gamma} = 100$. To demonstrate this, consider Tables 7.103-7.104, which have the same parameter values as Tables 7.1-7.2, except for $\hat{\gamma} = 10$ instead of 100.

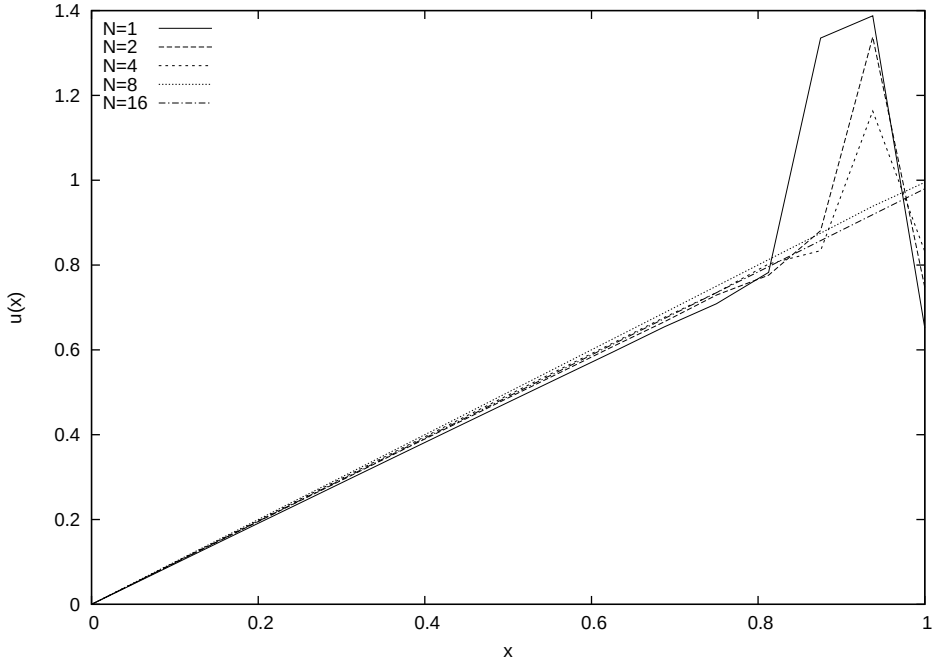

Figure 7.1: Numerical solutions $u^h(t = 1)$ for $Q = 4$ and $h = 1/16$. Here, exactness is given only for $N \geq 8$ time steps

h-convergence for $r = 1$

In the tables for h-convergence (Tables 7.45 to 7.88) the EOC for e_0 and e_{comb} is close to 2 and the EOC for e_1 is close to 1, as predicted by the theory.

The h-convergence for the nonlinear methods $Q = 4, 5$ is difficult to see in 2D. The time step needs to be small, compare the previous subsection (Subsection 7.4.2). Therefore many time steps are necessary, and also several iterations are needed at each time step to solve the nonlinear system. As the calculations take a lot longer than for the linearised methods, the tables have to be shorter. This is why the tables for 2D are too short to be able to confirm the convergence rate, see for example Table 7.51. However we refer the reader to the tables in 1D, Tables 7.95, 7.96, 7.99 and 7.100, where the convergence rate can be seen to agree with theory.

In some cases the choice of $\gamma \in \{\gamma_3, \gamma_1\}$ seems to have an influence, for example Table 7.51 (with γ_3) as compared to Table 7.61 (with γ_1). As mentioned at the beginning of this section, the value of Δ_a is chosen so that the size of the transition region of γ is comparable for both choices of γ. It is unclear why the small change in the shape of function γ causes these big differences in the errors.

h-convergence for $r \in \{2, 3, 4\}$

The tables for higher degree polynomials also indicate higher convergence rates. In the case of $r = 2$ (Tables 7.89 and 7.90), the EOC for e_0 and e_{comb} is close to 3 and the EOC for e_1 is close to 2, as predicted by theory. However for $r = 3$ and $r = 4$, (Tables 7.91 and 7.94) the orders of convergence from the theory cannot be clearly confirmed anymore. The tabulated values of the EOC are high but they are also varying a lot within one column. Also, when the error reaches an order of magnitude of 10^{-9} it does not decrease further. One reason for this could be that the linear systems are not solved accurately enough anymore, compare the values of the residual, which are also tabulated.

Table 7.1: k-convergence for $\sigma_D = 1$, $\gamma = \gamma_3$, $Q = 1$, $d = 2$, $r = 1$. Errors e_{comb}

N	k	e_{comb}	EOC	residual
1	$1.0000E+00$	$7.82E-01$	-	1.1297e-14
2	$5.0000E-01$	$6.62E-01$	0.2413	4.0489e-15
4	$2.5000E-01$	$3.60E-01$	0.8784	3.2784e-15
8	$1.2500E-01$	$1.77E-01$	1.0209	5.3797e-15
16	$6.2500E-02$	$9.84E-02$	0.8508	7.9843e-15
32	$3.1250E-02$	$7.11E-02$	0.4691	8.1450e-15
64	$1.5625E-02$	$4.88E-02$	0.5425	1.6305e-14
128	$7.8125E-03$	$2.24E-02$	1.1216	2.1683e-14
256	$3.9062E-03$	$7.22E-03$	1.6357	3.5527e-14
512	$1.9531E-03$	$1.95E-03$	1.8860	8.6441e-14
1024	$9.7656E-04$	$4.98E-04$	1.9709	1.8474e-13
2048	$4.8828E-04$	$1.25E-04$	1.9940	3.7813e-13

Table 7.2: k-convergence for $\sigma_D = 1$, $\gamma = \gamma_3$, $Q = 1$, $d = 2$, $r = 1$. Errors e_0 and e_1

N	k	e_0	EOC	e_1	EOC
1	$1.0000E+00$	$6.90E-01$	-	$9.20E-02$	-
2	$5.0000E-01$	$5.66E-01$	0.2875	$9.62E-02$	-0.0651
4	$2.5000E-01$	$3.04E-01$	0.8979	$5.65E-02$	0.7690
8	$1.2500E-01$	$1.49E-01$	1.0264	$2.84E-02$	0.9917
16	$6.2500E-02$	$8.25E-02$	0.8526	$1.58E-02$	0.8415
32	$3.1250E-02$	$5.86E-02$	0.4939	$1.25E-02$	0.3463
64	$1.5625E-02$	$3.96E-02$	0.5637	$9.15E-03$	0.4469
128	$7.8125E-03$	$1.81E-02$	1.1277	$4.28E-03$	1.0953
256	$3.9062E-03$	$5.83E-03$	1.6367	$1.38E-03$	1.6315
512	$1.9531E-03$	$1.58E-03$	1.8861	$3.74E-04$	1.8855
1024	$9.7656E-04$	$4.03E-04$	1.9709	$9.54E-05$	1.9708
2048	$4.8828E-04$	$1.01E-04$	1.9940	$2.39E-05$	1.9940

Table 7.3: k-convergence for $\sigma_D = 1$, $\gamma = \gamma_3$, $Q = 2$, $d = 2$, $r = 1$. Errors e_{comb}

N	k	e_{comb}	EOC	residual
1	$1.0000E+00$	$7.82E-01$	-	1.1297e-14
2	$5.0000E-01$	$6.94E-01$	0.1729	7.9643e-15
4	$2.5000E-01$	$4.73E-01$	0.5534	1.8519e-15
8	$1.2500E-01$	$3.13E-01$	0.5939	7.9982e-15
16	$6.2500E-02$	$2.24E-01$	0.4852	7.3384e-15
32	$3.1250E-02$	$1.87E-01$	0.2612	1.0166e-14
64	$1.5625E-02$	$1.56E-01$	0.2563	1.4349e-14
128	$7.8125E-03$	$8.69E-02$	0.8467	2.4678e-14
256	$3.9062E-03$	$3.14E-02$	1.4710	4.3221e-14
512	$1.9531E-03$	$8.89E-03$	1.8187	9.2371e-14
1024	$9.7656E-04$	$2.30E-03$	1.9481	1.4492e-13
2048	$4.8828E-04$	$5.81E-04$	1.9863	3.5951e-13

Table 7.4: k-convergence for $\sigma_D = 1$, $\gamma = \gamma_3$, $Q = 2$, $d = 2$, $r = 1$. Errors e_0 and e_1

N	k	e_0	EOC	e_1	EOC
1	$1.0000E+00$	$6.90E-01$	-	$9.20E-02$	-
2	$5.0000E-01$	$5.93E-01$	0.2194	$1.01E-01$	-0.1354
4	$2.5000E-01$	$3.98E-01$	0.5735	$7.44E-02$	0.4406
8	$1.2500E-01$	$2.63E-01$	0.6004	$5.05E-02$	0.5596
16	$6.2500E-02$	$1.88E-01$	0.4848	$3.60E-02$	0.4873
32	$3.1250E-02$	$1.55E-01$	0.2811	$3.22E-02$	0.1619
64	$1.5625E-02$	$1.27E-01$	0.2818	$2.92E-02$	0.1400
128	$7.8125E-03$	$7.03E-02$	0.8545	$1.66E-02$	0.8132
256	$3.9062E-03$	$2.53E-02$	1.4718	$6.02E-03$	1.4676
512	$1.9531E-03$	$7.19E-03$	1.8184	$1.70E-03$	1.8199
1024	$9.7656E-04$	$1.86E-03$	1.9480	$4.41E-04$	1.9487
2048	$4.8828E-04$	$4.70E-04$	1.9863	$1.11E-04$	1.9865

Table 7.5: k-convergence for $\sigma_D = 1$, $\gamma = \gamma_3$, $Q = 3$, $d = 2$, $r = 1$. Errors e_{comb}

N	k	e_{comb}	EOC	residual
1	$1.0000E+00$	$1.43E-01$	-	2.0063e-16
2	$5.0000E-01$	$1.28E-01$	0.1550	4.4755e-16
4	$2.5000E-01$	$1.10E-01$	0.2249	9.1045e-16
8	$1.2500E-01$	$4.02E-02$	1.4456	7.6919e-16
16	$6.2500E-02$	$8.59E-03$	2.2272	1.8310e-15
32	$3.1250E-02$	$1.90E-03$	2.1758	3.7682e-15
64	$1.5625E-02$	$4.51E-04$	2.0780	3.9721e-15
128	$7.8125E-03$	$1.08E-04$	2.0653	8.7023e-15
256	$3.9062E-03$	$2.60E-05$	2.0495	2.8422e-14
512	$1.9531E-03$	$6.35E-06$	2.0330	5.8593e-14
1024	$9.7656E-04$	$1.57E-06$	2.0193	1.1719e-13
2048	$4.8828E-04$	$3.89E-07$	2.0105	2.4117e-13

Table 7.6: k-convergence for $\sigma_D = 1$, $\gamma = \gamma_3$, $Q = 3$, $d = 2$, $r = 1$. Errors e_0 and e_1

N	k	e_0	EOC	e_1	EOC
1	$1.0000E+00$	$1.29E-01$	-	$1.32E-02$	-
2	$5.0000E-01$	$1.06E-01$	0.2839	$2.18E-02$	-0.7247
4	$2.5000E-01$	$8.89E-02$	0.2575	$2.07E-02$	0.0755
8	$1.2500E-01$	$3.32E-02$	1.4229	$7.07E-03$	1.5472
16	$6.2500E-02$	$7.29E-03$	2.1863	$1.31E-03$	2.4364
32	$3.1250E-02$	$1.61E-03$	2.1827	$2.97E-04$	2.1377
64	$1.5625E-02$	$3.82E-04$	2.0709	$6.84E-05$	2.1170
128	$7.8125E-03$	$9.16E-05$	2.0610	$1.61E-05$	2.0893
256	$3.9062E-03$	$2.22E-05$	2.0472	$3.85E-06$	2.0625
512	$1.9531E-03$	$5.42E-06$	2.0319	$9.37E-07$	2.0394
1024	$9.7656E-04$	$1.34E-06$	2.0188	$2.31E-07$	2.0225
2048	$4.8828E-04$	$3.32E-07$	2.0102	$5.72E-08$	2.0120

Table 7.7: k-convergence for $\sigma_D = 1$, $\gamma = \gamma_3$, $Q = 4$, $d = 2$, $r = 1$. Errors e_{comb}

N	k	e_{comb}	EOC	residual
1	$1.0000E+00$	$6.55E+01$	-	2.4364e-14
2	$5.0000E-01$	$4.04E+01$	0.6978	1.0476e-14
4	$2.5000E-01$	$2.08E+01$	0.9540	5.0502e-14
8	$1.2500E-01$	$4.84E+00$	2.1079	9.9794e-14
16	$6.2500E-02$	$6.35E+00$	-0.3923	7.3546e-14
32	$3.1250E-02$	$7.24E-01$	3.1326	9.9051e-14
64	$1.5625E-02$	$1.34E-03$	9.0809	8.7096e-14
128	$7.8125E-03$	$1.81E-04$	2.8881	9.5808e-14
256	$3.9062E-03$	$4.03E-05$	2.1634	9.3255e-14
512	$1.9531E-03$	$9.81E-06$	2.0382	1.2571e-13
1024	$9.7656E-04$	$2.44E-06$	2.0094	2.0740e-13
2048	$4.8828E-04$	$6.08E-07$	2.0023	4.7218e-13

Table 7.8: k-convergence for $\sigma_D = 1$, $\gamma = \gamma_3$, $Q = 4$, $d = 2$, $r = 1$. Errors e_0 and e_1

N	k	e_0	EOC	e_1	EOC
1	$1.0000E+00$	$5.63E+01$	-	$9.21E+00$	-
2	$5.0000E-01$	$3.40E+01$	0.7258	$6.35E+00$	0.5376
4	$2.5000E-01$	$1.74E+01$	0.9711	$3.48E+00$	0.8660
8	$1.2500E-01$	$4.04E+00$	2.1050	$8.00E-01$	2.1225
16	$6.2500E-02$	$5.20E+00$	-0.3665	$1.14E+00$	-0.5162
32	$3.1250E-02$	$5.91E-01$	3.1376	$1.32E-01$	3.1100
64	$1.5625E-02$	$1.08E-03$	9.0953	$2.55E-04$	9.0181
128	$7.8125E-03$	$1.47E-04$	2.8793	$3.36E-05$	2.9259
256	$3.9062E-03$	$3.28E-05$	2.1611	$7.45E-06$	2.1734
512	$1.9531E-03$	$8.00E-06$	2.0376	$1.81E-06$	2.0407
1024	$9.7656E-04$	$1.99E-06$	2.0092	$4.50E-07$	2.0100
2048	$4.8828E-04$	$4.96E-07$	2.0023	$1.12E-07$	2.0025

Table 7.9: k-convergence for $\sigma_D = 1$, $\gamma = \gamma_3$, $Q = 5$, $d = 2$, $r = 1$. Errors e_{comb}

N	k	e_{comb}	EOC	residual
1	$1.0000E+00$	$7.85E-01$	-	2.3367e-14
2	$5.0000E-01$	$7.34E-01$	0.0964	2.8935e-14
4	$2.5000E-01$	$6.33E-01$	0.2135	8.2022e-14
8	$1.2500E-01$	$9.20E+00$	-3.8601	9.3020e-14
16	$6.2500E-02$	$4.46E+00$	1.0436	8.5640e-14
32	$3.1250E-02$	$1.56E+00$	1.5175	9.6715e-14
64	$1.5625E-02$	$3.70E-01$	2.0758	9.3912e-14
128	$7.8125E-03$	$1.07E-01$	1.7836	9.1685e-14
256	$3.9062E-03$	$2.87E-02$	1.9024	9.8552e-14
512	$1.9531E-03$	$7.30E-03$	1.9759	1.1457e-13
1024	$9.7656E-04$	$1.83E-03$	1.9941	2.6088e-13
2048	$4.8828E-04$	$4.59E-04$	1.9985	4.1383e-13

Table 7.10: k-convergence for $\sigma_D = 1$, $\gamma = \gamma_3$, $Q = 5$, $d = 2$, $r = 1$. Errors e_0 and e_1

N	k	e_0	EOC	e_1	EOC
1	$1.0000E+00$	$6.93E-01$	-	$9.25E-02$	-
2	$5.0000E-01$	$6.27E-01$	0.1433	$1.07E-01$	-0.2134
4	$2.5000E-01$	$5.33E-01$	0.2343	$1.00E-01$	0.0975
8	$1.2500E-01$	$7.62E+00$	-3.8368	$1.58E+00$	-3.9780
16	$6.2500E-02$	$3.69E+00$	1.0450	$7.70E-01$	1.0368
32	$3.1250E-02$	$1.28E+00$	1.5272	$2.78E-01$	1.4718
64	$1.5625E-02$	$3.00E-01$	2.0962	$7.01E-02$	1.9851
128	$7.8125E-03$	$8.66E-02$	1.7900	$2.07E-02$	1.7569
256	$3.9062E-03$	$2.32E-02$	1.9005	$5.52E-03$	1.9103
512	$1.9531E-03$	$5.90E-03$	1.9751	$1.40E-03$	1.9792
1024	$9.7656E-04$	$1.48E-03$	1.9938	$3.51E-04$	1.9950
2048	$4.8828E-04$	$3.71E-04$	1.9985	$8.79E-05$	1.9988

Table 7.11: k-convergence for $\sigma_D = 1$, $\gamma = \gamma_1$, $Q = 1$, $d = 2$, $r = 1$. Errors e_{comb}

N	k	e_{comb}	EOC	residual
1	$1.0000E+00$	$7.89E-01$	-	1.8652e-14
2	$5.0000E-01$	$6.67E-01$	0.2420	1.0218e-14
4	$2.5000E-01$	$3.61E-01$	0.8849	7.9596e-15
8	$1.2500E-01$	$1.75E-01$	1.0464	8.0190e-15
16	$6.2500E-02$	$9.98E-02$	0.8097	5.4265e-15
32	$3.1250E-02$	$7.24E-02$	0.4626	7.9938e-15
64	$1.5625E-02$	$4.99E-02$	0.5365	9.3996e-15
128	$7.8125E-03$	$2.29E-02$	1.1232	2.1390e-14
256	$3.9062E-03$	$7.36E-03$	1.6387	4.3945e-14
512	$1.9531E-03$	$1.99E-03$	1.8876	9.6644e-14
1024	$9.7656E-04$	$5.07E-04$	1.9714	1.6078e-13
2048	$4.8828E-04$	$1.27E-04$	1.9941	4.8899e-13

Table 7.12: k-convergence for $\sigma_D = 1$, $\gamma = \gamma_1$, $Q = 1$, $d = 2$, $r = 1$. Errors e_0 and e_1

N	k	e_0	EOC	e_1	EOC
1	$1.0000E+00$	$6.96E-01$	-	$9.31E-02$	-
2	$5.0000E-01$	$5.70E-01$	0.2881	$9.72E-02$	-0.0626
4	$2.5000E-01$	$3.04E-01$	0.9041	$5.67E-02$	0.7769
8	$1.2500E-01$	$1.47E-01$	1.0514	$2.80E-02$	1.0201
16	$6.2500E-02$	$8.37E-02$	0.8113	$1.60E-02$	0.8015
32	$3.1250E-02$	$5.97E-02$	0.4876	$1.27E-02$	0.3385
64	$1.5625E-02$	$4.06E-02$	0.5575	$9.34E-03$	0.4415
128	$7.8125E-03$	$1.85E-02$	1.1293	$4.37E-03$	1.0968
256	$3.9062E-03$	$5.95E-03$	1.6398	$1.41E-03$	1.6342
512	$1.9531E-03$	$1.61E-03$	1.8878	$3.81E-04$	1.8870
1024	$9.7656E-04$	$4.10E-04$	1.9714	$9.70E-05$	1.9713
2048	$4.8828E-04$	$1.03E-04$	1.9942	$2.44E-05$	1.9941

Table 7.13: k-convergence for $\sigma_D = 1$, $\gamma = \gamma_1$, $Q = 2$, $d = 2$, $r = 1$. Errors e_{comb}

N	k	e_{comb}	EOC	residual
1	$1.0000E+00$	$7.89E-01$	-	1.8652e-14
2	$5.0000E-01$	$6.99E-01$	0.1751	5.1097e-15
4	$2.5000E-01$	$4.73E-01$	0.5614	7.8221e-15
8	$1.2500E-01$	$3.13E-01$	0.5950	8.1796e-15
16	$6.2500E-02$	$2.28E-01$	0.4612	7.3543e-15
32	$3.1250E-02$	$1.93E-01$	0.2379	1.0178e-14
64	$1.5625E-02$	$1.62E-01$	0.2555	1.4862e-14
128	$7.8125E-03$	$8.86E-02$	0.8673	2.4096e-14
256	$3.9062E-03$	$3.18E-02$	1.4792	4.5497e-14
512	$1.9531E-03$	$9.00E-03$	1.8216	9.6383e-14
1024	$9.7656E-04$	$2.33E-03$	1.9491	1.8474e-13
2048	$4.8828E-04$	$5.88E-04$	1.9866	3.4106e-13

Table 7.14: k-convergence for $\sigma_D = 1$, $\gamma = \gamma_1$, $Q = 2$, $d = 2$, $r = 1$. Errors e_0 and e_1

N	k	e_0	EOC	e_1	EOC
1	$1.0000E+00$	$6.96E-01$	-	$9.31E-02$	-
2	$5.0000E-01$	$5.97E-01$	0.2215	$1.02E-01$	-0.1313
4	$2.5000E-01$	$3.99E-01$	0.5814	$7.46E-02$	0.4496
8	$1.2500E-01$	$2.63E-01$	0.6006	$5.04E-02$	0.5653
16	$6.2500E-02$	$1.91E-01$	0.4612	$3.66E-02$	0.4612
32	$3.1250E-02$	$1.60E-01$	0.2581	$3.33E-02$	0.1370
64	$1.5625E-02$	$1.32E-01$	0.2804	$3.02E-02$	0.1419
128	$7.8125E-03$	$7.17E-02$	0.8750	$1.69E-02$	0.8343
256	$3.9062E-03$	$2.57E-02$	1.4802	$6.09E-03$	1.4753
512	$1.9531E-03$	$7.27E-03$	1.8213	$1.72E-03$	1.8225
1024	$9.7656E-04$	$1.88E-03$	1.9490	$4.46E-04$	1.9496
2048	$4.8828E-04$	$4.75E-04$	1.9866	$1.12E-04$	1.9867

Table 7.15: k-convergence for $\sigma_D = 1$, $\gamma = \gamma_1$, $Q = 3$, $d = 2$, $r = 1$. Errors e_{comb}

N	k	e_{comb}	EOC	residual
1	$1.0000E+00$	$1.62E-01$	-	2.8844e-16
2	$5.0000E-01$	$1.45E-01$	0.1547	4.4409e-16
4	$2.5000E-01$	$1.06E-01$	0.4538	8.8991e-16
8	$1.2500E-01$	$3.56E-02$	1.5722	6.7532e-16
16	$6.2500E-02$	$8.46E-03$	2.0751	1.0878e-15
32	$3.1250E-02$	$1.97E-03$	2.1013	2.7013e-15
64	$1.5625E-02$	$4.76E-04$	2.0511	8.1403e-15
128	$7.8125E-03$	$1.16E-04$	2.0348	1.4211e-14
256	$3.9062E-03$	$2.84E-05$	2.0307	2.8422e-14
512	$1.9531E-03$	$6.99E-06$	2.0228	6.0292e-14
1024	$9.7656E-04$	$1.73E-06$	2.0141	1.2058e-13
2048	$4.8828E-04$	$4.30E-07$	2.0079	2.4117e-13

Table 7.16: k-convergence for $\sigma_D = 1$, $\gamma = \gamma_1$, $Q = 3$, $d = 2$, $r = 1$. Errors e_0 and e_1

N	k	e_0	EOC	e_1	EOC
1	$1.0000E+00$	$1.45E-01$	-	$1.65E-02$	-
2	$5.0000E-01$	$1.20E-01$	0.2729	$2.51E-02$	-0.6025
4	$2.5000E-01$	$8.60E-02$	0.4811	$2.00E-02$	0.3303
8	$1.2500E-01$	$2.93E-02$	1.5519	$6.31E-03$	1.6629
16	$6.2500E-02$	$7.07E-03$	2.0530	$1.39E-03$	2.1823
32	$3.1250E-02$	$1.66E-03$	2.0914	$3.13E-04$	2.1523
64	$1.5625E-02$	$4.02E-04$	2.0457	$7.39E-05$	2.0802
128	$7.8125E-03$	$9.83E-05$	2.0319	$1.78E-05$	2.0502
256	$3.9062E-03$	$2.41E-05$	2.0295	$4.35E-06$	2.0374
512	$1.9531E-03$	$5.93E-06$	2.0223	$1.07E-06$	2.0255
1024	$9.7656E-04$	$1.47E-06$	2.0139	$2.64E-07$	2.0152
2048	$4.8828E-04$	$3.65E-07$	2.0078	$6.57E-08$	2.0084

Table 7.17: k-convergence for $\sigma_D = 1$, $\gamma = \gamma_1$, $Q = 4$, $d = 2$, $r = 1$. Errors e_{comb}

N	k	e_{comb}	EOC	residual
1	$1.0000E+00$	$6.54E+01$	-	3.6285e-14
2	$5.0000E-01$	$4.03E+01$	0.6977	5.0967e-14
4	$2.5000E-01$	$2.08E+01$	0.9560	5.9131e-14
8	$1.2500E-01$	$1.11E+01$	0.9066	9.3072e-14
16	$6.2500E-02$	$6.30E+00$	0.8149	9.7330e-14
32	$3.1250E-02$	$6.85E-01$	3.2022	9.4483e-14
64	$1.5625E-02$	$1.25E-03$	9.1002	9.9891e-14
128	$7.8125E-03$	$1.75E-04$	2.8360	9.5808e-14
256	$3.9062E-03$	$3.92E-05$	2.1558	9.6544e-14
512	$1.9531E-03$	$9.56E-06$	2.0365	1.1995e-13
1024	$9.7656E-04$	$2.38E-06$	2.0090	1.9329e-13
2048	$4.8828E-04$	$5.93E-07$	2.0022	3.9280e-13

Table 7.18: k-convergence for $\sigma_D = 1$, $\gamma = \gamma_1$, $Q = 4$, $d = 2$, $r = 1$. Errors e_0 and e_1

N	k	e_0	EOC	e_1	EOC
1	$1.0000E+00$	$5.62E+01$	-	$9.20E+00$	-
2	$5.0000E-01$	$3.40E+01$	0.7257	$6.34E+00$	0.5375
4	$2.5000E-01$	$1.73E+01$	0.9730	$3.47E+00$	0.8681
8	$1.2500E-01$	$9.14E+00$	0.9213	$1.95E+00$	0.8351
16	$6.2500E-02$	$5.17E+00$	0.8228	$1.13E+00$	0.7785
32	$3.1250E-02$	$5.59E-01$	3.2085	$1.26E-01$	3.1738
64	$1.5625E-02$	$1.01E-03$	9.1129	$2.38E-04$	9.0451
128	$7.8125E-03$	$1.42E-04$	2.8275	$3.25E-05$	2.8730
256	$3.9062E-03$	$3.20E-05$	2.1536	$7.24E-06$	2.1656
512	$1.9531E-03$	$7.80E-06$	2.0359	$1.76E-06$	2.0389
1024	$9.7656E-04$	$1.94E-06$	2.0088	$4.38E-07$	2.0096
2048	$4.8828E-04$	$4.84E-07$	2.0022	$1.09E-07$	2.0024

Table 7.19: k-convergence for $\sigma_D = 1$, $\gamma = \gamma_1$, $Q = 5$, $d = 2$, $r = 1$. Errors e_{comb}

N	k	e_{comb}	EOC	residual
1	$1.0000E+00$	$7.92E-01$	-	9.5724e-15
2	$5.0000E-01$	$7.41E-01$	0.0961	3.5352e-14
4	$2.5000E-01$	$6.39E-01$	0.2140	5.5639e-14
8	$1.2500E-01$	$9.13E+00$	-3.8368	8.3718e-14
16	$6.2500E-02$	$4.44E+00$	1.0409	8.6743e-14
32	$3.1250E-02$	$1.61E+00$	1.4642	9.9833e-14
64	$1.5625E-02$	$3.88E-01$	2.0501	9.1357e-14
128	$7.8125E-03$	$1.09E-01$	1.8332	9.6791e-14
256	$3.9062E-03$	$2.90E-02$	1.9101	9.5495e-14
512	$1.9531E-03$	$7.37E-03$	1.9761	9.9476e-14
1024	$9.7656E-04$	$1.85E-03$	1.9941	2.1834e-13
2048	$4.8828E-04$	$4.63E-04$	1.9985	4.4575e-13

Table 7.20: k-convergence for $\sigma_D = 1$, $\gamma = \gamma_1$, $Q = 5$, $d = 2$, $r = 1$. Errors e_0 and e_1

N	k	e_0	EOC	e_1	EOC
1	$1.0000E+00$	$6.98E-01$	-	$9.36E-02$	-
2	$5.0000E-01$	$6.32E-01$	0.1428	$1.08E-01$	-0.2116
4	$2.5000E-01$	$5.37E-01$	0.2348	$1.01E-01$	0.0985
8	$1.2500E-01$	$7.56E+00$	-3.8142	$1.57E+00$	-3.9514
16	$6.2500E-02$	$3.67E+00$	1.0425	$7.65E-01$	1.0335
32	$3.1250E-02$	$1.32E+00$	1.4736	$2.86E-01$	1.4199
64	$1.5625E-02$	$3.15E-01$	2.0707	$7.36E-02$	1.9586
128	$7.8125E-03$	$8.79E-02$	1.8392	$2.10E-02$	1.8082
256	$3.9062E-03$	$2.34E-02$	1.9085	$5.56E-03$	1.9169
512	$1.9531E-03$	$5.96E-03$	1.9753	$1.41E-03$	1.9792
1024	$9.7656E-04$	$1.50E-03$	1.9939	$3.54E-04$	1.9950
2048	$4.8828E-04$	$3.74E-04$	1.9985	$8.86E-05$	1.9988

Table 7.21: k-convergence for $\sigma_D = 0$, $\gamma = \gamma_3$, $Q = 1$, $d = 2$, $r = 1$. Errors e_{comb}

N	k	e_{comb}	EOC	residual
1	$1.0000E+00$	$1.85E+00$	-	8.8135e-15
2	$5.0000E-01$	$1.75E+00$	0.0814	1.1828e-15
4	$2.5000E-01$	$2.71E-01$	2.6882	5.8400e-15
8	$1.2500E-01$	$6.09E-02$	2.1540	4.4312e-15
16	$6.2500E-02$	$3.04E-02$	1.0018	6.1864e-15
32	$3.1250E-02$	$2.61E-02$	0.2191	8.2488e-15
64	$1.5625E-02$	$1.87E-02$	0.4848	1.4862e-14
128	$7.8125E-03$	$8.48E-03$	1.1380	2.1610e-14
256	$3.9062E-03$	$2.71E-03$	1.6453	4.8322e-14
512	$1.9531E-03$	$7.35E-04$	1.8835	1.1804e-13
1024	$9.7656E-04$	$1.88E-04$	1.9664	2.4367e-13
2048	$4.8828E-04$	$4.73E-05$	1.9908	4.3010e-13

Table 7.22: k-convergence for $\sigma_D = 0$, $\gamma = \gamma_3$, $Q = 1$, $d = 2$, $r = 1$. Errors e_0 and e_1

N	k	e_0	EOC	e_1	EOC
1	$1.0000E+00$	$1.12E+00$	-	$7.28E-01$	-
2	$5.0000E-01$	$1.12E+00$	-0.0045	$6.23E-01$	0.2245
4	$2.5000E-01$	$2.01E-01$	2.4791	$6.94E-02$	3.1649
8	$1.2500E-01$	$4.63E-02$	2.1209	$1.46E-02$	2.2546
16	$6.2500E-02$	$2.49E-02$	0.8984	$5.55E-03$	1.3912
32	$3.1250E-02$	$2.13E-02$	0.2225	$4.82E-03$	0.2040
64	$1.5625E-02$	$1.51E-02$	0.5003	$3.60E-03$	0.4180
128	$7.8125E-03$	$6.82E-03$	1.1428	$1.66E-03$	1.1182
256	$3.9062E-03$	$2.18E-03$	1.6462	$5.32E-04$	1.6414
512	$1.9531E-03$	$5.90E-04$	1.8837	$1.44E-04$	1.8828
1024	$9.7656E-04$	$1.51E-04$	1.9664	$3.69E-05$	1.9662
2048	$4.8828E-04$	$3.80E-05$	1.9909	$9.29E-06$	1.9908

Table 7.23: k-convergence for $\sigma_D = 0$, $\gamma = \gamma_3$, $Q = 2$, $d = 2$, $r = 1$. Errors e_{comb}

N	k	e_{comb}	EOC	residual
1	$1.0000E+00$	$1.85E+00$	-	8.8135e-15
2	$5.0000E-01$	$1.70E+00$	0.1175	6.5107e-15
4	$2.5000E-01$	$2.70E-01$	2.6554	1.6513e-14
8	$1.2500E-01$	$1.32E-01$	1.0344	8.2911e-15
16	$6.2500E-02$	$6.23E-02$	1.0831	5.6567e-15
32	$3.1250E-02$	$3.73E-02$	0.7402	1.2016e-14
64	$1.5625E-02$	$2.56E-02$	0.5429	2.2749e-14
128	$7.8125E-03$	$1.23E-02$	1.0534	2.3297e-14
256	$3.9062E-03$	$4.05E-03$	1.6047	4.4373e-14
512	$1.9531E-03$	$1.11E-03$	1.8710	8.5265e-14
1024	$9.7656E-04$	$2.84E-04$	1.9635	1.8254e-13
2048	$4.8828E-04$	$7.15E-05$	1.9904	4.2252e-13

Table 7.24: k-convergence for $\sigma_D = 0$, $\gamma = \gamma_3$, $Q = 2$, $d = 2$, $r = 1$. Errors e_0 and e_1

N	k	e_0	EOC	e_1	EOC
1	$1.0000E+00$	$1.12E+00$	-	$7.28E-01$	-
2	$5.0000E-01$	$1.10E+00$	0.0289	$6.05E-01$	0.2655
4	$2.5000E-01$	$1.80E-01$	2.6049	$8.99E-02$	2.7517
8	$1.2500E-01$	$8.81E-02$	1.0346	$4.39E-02$	1.0342
16	$6.2500E-02$	$4.71E-02$	0.9024	$1.52E-02$	1.5329
32	$3.1250E-02$	$2.99E-02$	0.6548	$7.36E-03$	1.0434
64	$1.5625E-02$	$2.06E-02$	0.5384	$4.98E-03$	0.5616
128	$7.8125E-03$	$9.91E-03$	1.0568	$2.42E-03$	1.0397
256	$3.9062E-03$	$3.26E-03$	1.6053	$7.99E-04$	1.6020
512	$1.9531E-03$	$8.90E-04$	1.8711	$2.18E-04$	1.8706
1024	$9.7656E-04$	$2.28E-04$	1.9635	$5.60E-05$	1.9634
2048	$4.8828E-04$	$5.74E-05$	1.9904	$1.41E-05$	1.9904

Table 7.25: k-convergence for $\sigma_D = 0$, $\gamma = \gamma_3$, $Q = 3$, $d = 2$, $r = 1$. Errors e_{comb}

N	k	e_{comb}	EOC	residual
1	$1.0000E+00$	$1.51E-01$	-	1.8619e-16
2	$5.0000E-01$	$1.28E-01$	0.2366	1.6653e-16
4	$2.5000E-01$	$1.10E-01$	0.2159	4.4409e-16
8	$1.2500E-01$	$4.13E-02$	1.4136	5.4390e-16
16	$6.2500E-02$	$8.99E-03$	2.2000	1.0878e-15
32	$3.1250E-02$	$1.96E-03$	2.1986	3.5527e-15
64	$1.5625E-02$	$4.74E-04$	2.0484	7.3241e-15
128	$7.8125E-03$	$1.16E-04$	2.0325	1.0049e-14
256	$3.9062E-03$	$2.83E-05$	2.0338	3.0146e-14
512	$1.9531E-03$	$6.93E-06$	2.0281	6.0292e-14
1024	$9.7656E-04$	$1.71E-06$	2.0182	1.1719e-13
2048	$4.8828E-04$	$4.25E-07$	2.0104	2.6049e-13

Table 7.26: k-convergence for $\sigma_D = 0$, $\gamma = \gamma_3$, $Q = 3$, $d = 2$, $r = 1$. Errors e_0 and e_1

N	k	e_0	EOC	e_1	EOC
1	$1.0000E+00$	$1.23E-01$	-	$2.73E-02$	-
2	$5.0000E-01$	$1.05E-01$	0.2285	$2.26E-02$	0.2737
4	$2.5000E-01$	$9.20E-02$	0.1939	$1.81E-02$	0.3234
8	$1.2500E-01$	$3.55E-02$	1.3744	$5.83E-03$	1.6319
16	$6.2500E-02$	$7.93E-03$	2.1618	$1.06E-03$	2.4577
32	$3.1250E-02$	$1.71E-03$	2.2155	$2.51E-04$	2.0783
64	$1.5625E-02$	$4.12E-04$	2.0524	$6.19E-05$	2.0215
128	$7.8125E-03$	$1.01E-04$	2.0341	$1.52E-05$	2.0216
256	$3.9062E-03$	$2.46E-05$	2.0339	$3.72E-06$	2.0333
512	$1.9531E-03$	$6.02E-06$	2.0276	$9.11E-07$	2.0308
1024	$9.7656E-04$	$1.49E-06$	2.0178	$2.24E-07$	2.0208
2048	$4.8828E-04$	$3.69E-07$	2.0101	$5.57E-08$	2.0120

Table 7.27: k-convergence for $\sigma_D = 0$, $\gamma = \gamma_3$, $Q = 4$, $d = 2$, $r = 1$. Errors e_{comb}

N	k	e_{comb}	EOC	residual
1	$1.0000E+00$	$8.81E+01$	-	2.5362e-14
2	$5.0000E-01$	$5.02E+01$	0.8096	2.3785e-14
4	$2.5000E-01$	$1.33E+01$	1.9201	9.2404e-14
8	$1.2500E-01$	$6.29E+00$	1.0766	9.9913e-14
16	$6.2500E-02$	$8.31E+00$	-0.4011	9.2849e-14
32	$3.1250E-02$	$3.95E+00$	1.0736	9.7481e-14
64	$1.5625E-02$	$7.98E-03$	8.9502	9.7845e-14
128	$7.8125E-03$	$9.27E-04$	3.1064	9.8091e-14
256	$3.9062E-03$	$2.01E-04$	2.2053	9.9540e-14
512	$1.9531E-03$	$4.86E-05$	2.0485	1.1847e-13
1024	$9.7656E-04$	$1.20E-05$	2.0120	2.4491e-13
2048	$4.8828E-04$	$3.01E-06$	2.0030	4.8567e-13

Table 7.28: k-convergence for $\sigma_D = 0$, $\gamma = \gamma_3$, $Q = 4$, $d = 2$, $r = 1$. Errors e_0 and e_1

N	k	e_0	EOC	e_1	EOC
1	$1.0000E+00$	$6.44E+01$	-	$2.37E+01$	-
2	$5.0000E-01$	$3.80E+01$	0.7601	$1.22E+01$	0.9533
4	$2.5000E-01$	$1.08E+01$	1.8100	$2.43E+00$	2.3284
8	$1.2500E-01$	$5.11E+00$	1.0865	$1.19E+00$	1.0335
16	$6.2500E-02$	$6.67E+00$	-0.3857	$1.64E+00$	-0.4655
32	$3.1250E-02$	$3.13E+00$	1.0906	$8.18E-01$	1.0062
64	$1.5625E-02$	$6.39E-03$	8.9376	$1.60E-03$	8.9995
128	$7.8125E-03$	$7.44E-04$	3.1022	$1.83E-04$	3.1234
256	$3.9062E-03$	$1.61E-04$	2.2044	$3.97E-05$	2.2089
512	$1.9531E-03$	$3.90E-05$	2.0483	$9.58E-06$	2.0494
1024	$9.7656E-04$	$9.67E-06$	2.0119	$2.37E-06$	2.0122
2048	$4.8828E-04$	$2.41E-06$	2.0030	$5.92E-07$	2.0030

Table 7.29: k-convergence for $\sigma_D = 0$, $\gamma = \gamma_3$, $Q = 5$, $d = 2$, $r = 1$. Errors e_{comb}

N	k	e_{comb}	EOC	residual
1	$1.0000E+00$	$5.21E+01$	-	4.9319e-01
2	$5.0000E-01$	$2.79E+01$	0.9014	4.8206e-14
4	$2.5000E-01$	$1.78E+01$	0.6492	8.5229e-14
8	$1.2500E-01$	$1.14E+01$	0.6452	8.5734e-14
16	$6.2500E-02$	$6.02E+00$	0.9165	9.3595e-14
32	$3.1250E-02$	$2.47E+00$	1.2848	9.8584e-14
64	$1.5625E-02$	$5.34E-02$	5.5337	9.9305e-14
128	$7.8125E-03$	$8.24E-03$	2.6952	9.5263e-14
256	$3.9062E-03$	$1.90E-03$	2.1177	9.9856e-14
512	$1.9531E-03$	$4.66E-04$	2.0270	1.1441e-13
1024	$9.7656E-04$	$1.16E-04$	2.0067	3.0479e-13
2048	$4.8828E-04$	$2.89E-05$	2.0017	4.9718e-13

Table 7.30: k-convergence for $\sigma_D = 0$, $\gamma = \gamma_3$, $Q = 5$, $d = 2$, $r = 1$. Errors e_0 and e_1

N	k	e_0	EOC	e_1	EOC
1	$1.0000E+00$	$4.07E+01$	-	$1.13E+01$	-
2	$5.0000E-01$	$2.30E+01$	0.8240	$4.86E+00$	1.2201
4	$2.5000E-01$	$1.46E+01$	0.6542	$3.15E+00$	0.6259
8	$1.2500E-01$	$9.26E+00$	0.6587	$2.10E+00$	0.5839
16	$6.2500E-02$	$4.84E+00$	0.9353	$1.18E+00$	0.8363
32	$3.1250E-02$	$1.97E+00$	1.3010	$5.06E-01$	1.2202
64	$1.5625E-02$	$4.27E-02$	5.5264	$1.07E-02$	5.5624
128	$7.8125E-03$	$6.62E-03$	2.6886	$1.62E-03$	2.7217
256	$3.9062E-03$	$1.53E-03$	2.1162	$3.72E-04$	2.1240
512	$1.9531E-03$	$3.75E-04$	2.0266	$9.12E-05$	2.0286
1024	$9.7656E-04$	$9.32E-05$	2.0066	$2.27E-05$	2.0070
2048	$4.8828E-04$	$2.33E-05$	2.0016	$5.66E-06$	2.0018

Table 7.31: k-convergence for $\sigma_D = 0$, $\gamma = \gamma_1$, $Q = 1$, $d = 2$, $r = 1$. Errors e_{comb}

N	k	e_{comb}	EOC	residual
1	$1.0000E+00$	$1.79E+00$	-	5.8946e-15
2	$5.0000E-01$	$1.64E+00$	0.1198	1.4496e-14
4	$2.5000E-01$	$2.49E-01$	2.7225	9.4225e-15
8	$1.2500E-01$	$5.82E-02$	2.0973	4.4464e-15
16	$6.2500E-02$	$3.17E-02$	0.8765	1.1248e-14
32	$3.1250E-02$	$2.71E-02$	0.2245	8.2053e-15
64	$1.5625E-02$	$1.97E-02$	0.4642	1.6012e-14
128	$7.8125E-03$	$8.98E-03$	1.1310	3.0250e-14
256	$3.9062E-03$	$2.87E-03$	1.6453	4.3801e-14
512	$1.9531E-03$	$7.78E-04$	1.8840	1.2888e-13
1024	$9.7656E-04$	$1.99E-04$	1.9665	2.1645e-13
2048	$4.8828E-04$	$5.01E-05$	1.9908	4.1577e-13

Table 7.32: k-convergence for $\sigma_D = 0$, $\gamma = \gamma_1$, $Q = 1$, $d = 2$, $r = 1$. Errors e_0 and e_1

N	k	e_0	EOC	e_1	EOC
1	$1.0000E+00$	$1.09E+00$	-	$6.95E-01$	-
2	$5.0000E-01$	$1.07E+00$	0.0327	$5.77E-01$	0.2683
4	$2.5000E-01$	$1.82E-01$	2.5485	$6.66E-02$	3.1133
8	$1.2500E-01$	$4.41E-02$	2.0487	$1.41E-02$	2.2396
16	$6.2500E-02$	$2.60E-02$	0.7635	$5.73E-03$	1.3005
32	$3.1250E-02$	$2.21E-02$	0.2307	$5.00E-03$	0.1966
64	$1.5625E-02$	$1.59E-02$	0.4795	$3.79E-03$	0.3984
128	$7.8125E-03$	$7.23E-03$	1.1358	$1.76E-03$	1.1112
256	$3.9062E-03$	$2.31E-03$	1.6463	$5.63E-04$	1.6413
512	$1.9531E-03$	$6.25E-04$	1.8842	$1.53E-04$	1.8832
1024	$9.7656E-04$	$1.60E-04$	1.9665	$3.90E-05$	1.9663
2048	$4.8828E-04$	$4.03E-05$	1.9908	$9.82E-06$	1.9907

Table 7.33: k-convergence for $\sigma_D = 0$, $\gamma = \gamma_1$, $Q = 2$, $d = 2$, $r = 1$. Errors e_{comb}

N	k	e_{comb}	EOC	residual
1	$1.0000E+00$	$1.79E+00$	-	5.8946e-15
2	$5.0000E-01$	$1.62E+00$	0.1427	1.8609e-14
4	$2.5000E-01$	$2.67E-01$	2.6018	3.6972e-15
8	$1.2500E-01$	$1.24E-01$	1.1000	8.2246e-15
16	$6.2500E-02$	$6.18E-02$	1.0095	5.3717e-15
32	$3.1250E-02$	$3.54E-02$	0.8024	5.5470e-15
64	$1.5625E-02$	$2.31E-02$	0.6169	1.3987e-14
128	$7.8125E-03$	$1.11E-02$	1.0603	2.9564e-14
256	$3.9062E-03$	$3.64E-03$	1.6053	5.5267e-14
512	$1.9531E-03$	$9.94E-04$	1.8721	1.2348e-13
1024	$9.7656E-04$	$2.55E-04$	1.9640	2.1738e-13
2048	$4.8828E-04$	$6.41E-05$	1.9906	4.5563e-13

Table 7.34: k-convergence for $\sigma_D = 0$, $\gamma = \gamma_1$, $Q = 2$, $d = 2$, $r = 1$. Errors e_0 and e_1

N	k	e_0	EOC	e_1	EOC
1	$1.0000E+00$	$1.09E+00$	-	$6.95E-01$	-
2	$5.0000E-01$	$1.05E+00$	0.0532	$5.66E-01$	0.2957
4	$2.5000E-01$	$1.82E-01$	2.5304	$8.44E-02$	2.7447
8	$1.2500E-01$	$8.25E-02$	1.1431	$4.19E-02$	1.0110
16	$6.2500E-02$	$4.72E-02$	0.8040	$1.45E-02$	1.5270
32	$3.1250E-02$	$2.85E-02$	0.7276	$6.90E-03$	1.0760
64	$1.5625E-02$	$1.86E-02$	0.6162	$4.49E-03$	0.6195
128	$7.8125E-03$	$8.90E-03$	1.0640	$2.18E-03$	1.0448
256	$3.9062E-03$	$2.92E-03$	1.6060	$7.17E-04$	1.6022
512	$1.9531E-03$	$7.99E-04$	1.8722	$1.96E-04$	1.8715
1024	$9.7656E-04$	$2.05E-04$	1.9640	$5.02E-05$	1.9638
2048	$4.8828E-04$	$5.15E-05$	1.9906	$1.26E-05$	1.9906

Table 7.35: k-convergence for $\sigma_D = 0$, $\gamma = \gamma_1$, $Q = 3$, $d = 2$, $r = 1$. Errors e_{comb}

N	k	e_{comb}	EOC	residual
1	$1.0000E+00$	$1.69E-01$	-	5.7169e-16
2	$5.0000E-01$	$1.45E-01$	0.2270	3.6822e-16
4	$2.5000E-01$	$1.06E-01$	0.4441	5.6610e-16
8	$1.2500E-01$	$3.63E-02$	1.5524	6.2804e-16
16	$6.2500E-02$	$8.56E-03$	2.0831	1.8310e-15
32	$3.1250E-02$	$2.01E-03$	2.0881	3.7682e-15
64	$1.5625E-02$	$5.00E-04$	2.0098	7.3241e-15
128	$7.8125E-03$	$1.25E-04$	2.0022	1.4648e-14
256	$3.9062E-03$	$3.09E-05$	2.0161	2.8422e-14
512	$1.9531E-03$	$7.62E-06$	2.0180	5.8593e-14
1024	$9.7656E-04$	$1.89E-06$	2.0128	1.2058e-13
2048	$4.8828E-04$	$4.70E-07$	2.0076	2.5421e-13

Table 7.36: k-convergence for $\sigma_D = 0$, $\gamma = \gamma_1$, $Q = 3$, $d = 2$, $r = 1$. Errors e_0 and e_1

N	k	e_0	EOC	e_1	EOC
1	$1.0000E+00$	$1.38E-01$	-	$3.12E-02$	-
2	$5.0000E-01$	$1.19E-01$	0.2167	$2.58E-02$	0.2732
4	$2.5000E-01$	$8.90E-02$	0.4185	$1.74E-02$	0.5681
8	$1.2500E-01$	$3.11E-02$	1.5169	$5.19E-03$	1.7489
16	$6.2500E-02$	$7.42E-03$	2.0675	$1.14E-03$	2.1802
32	$3.1250E-02$	$1.74E-03$	2.0920	$2.74E-04$	2.0633
64	$1.5625E-02$	$4.31E-04$	2.0151	$6.96E-05$	1.9766
128	$7.8125E-03$	$1.07E-04$	2.0050	$1.76E-05$	1.9852
256	$3.9062E-03$	$2.65E-05$	2.0170	$4.36E-06$	2.0103
512	$1.9531E-03$	$6.54E-06$	2.0182	$1.08E-06$	2.0168
1024	$9.7656E-04$	$1.62E-06$	2.0128	$2.67E-07$	2.0128
2048	$4.8828E-04$	$4.03E-07$	2.0075	$6.64E-08$	2.0077

Table 7.37: k-convergence for $\sigma_D = 0$, $\gamma = \gamma_1$, $Q = 4$, $d = 2$, $r = 1$. Errors e_{comb}

N	k	e_{comb}	EOC	residual
1	$1.0000E+00$	$8.75E+01$	-	8.6121e-15
2	$5.0000E-01$	$4.04E+01$	1.1170	7.0380e-14
4	$2.5000E-01$	$1.29E+01$	1.6431	8.8616e-14
8	$1.2500E-01$	$6.25E+00$	1.0476	8.8402e-14
16	$6.2500E-02$	$8.27E+00$	-0.4037	9.7702e-14
32	$3.1250E-02$	$3.86E+00$	1.1001	9.6682e-14
64	$1.5625E-02$	$7.56E-03$	8.9955	9.9575e-14
128	$7.8125E-03$	$9.18E-04$	3.0414	9.5742e-14
256	$3.9062E-03$	$2.00E-04$	2.1970	9.8983e-14
512	$1.9531E-03$	$4.85E-05$	2.0467	9.9982e-14
1024	$9.7656E-04$	$1.20E-05$	2.0115	2.6280e-13
2048	$4.8828E-04$	$3.00E-06$	2.0029	4.4031e-13

Table 7.38: k-convergence for $\sigma_D = 0$, $\gamma = \gamma_1$, $Q = 4$, $d = 2$, $r = 1$. Errors e_0 and e_1

N	k	e_0	EOC	e_1	EOC
1	$1.0000E+00$	$6.42E+01$	-	$2.34E+01$	-
2	$5.0000E-01$	$3.40E+01$	0.9153	$6.33E+00$	1.8838
4	$2.5000E-01$	$1.06E+01$	1.6842	$2.33E+00$	1.4400
8	$1.2500E-01$	$5.07E+00$	1.0614	$1.18E+00$	0.9869
16	$6.2500E-02$	$6.64E+00$	-0.3883	$1.63E+00$	-0.4683
32	$3.1250E-02$	$3.06E+00$	1.1173	$7.96E-01$	1.0319
64	$1.5625E-02$	$6.05E-03$	8.9833	$1.51E-03$	9.0436
128	$7.8125E-03$	$7.37E-04$	3.0374	$1.81E-04$	3.0577
256	$3.9062E-03$	$1.61E-04$	2.1961	$3.94E-05$	2.2005
512	$1.9531E-03$	$3.89E-05$	2.0465	$9.54E-06$	2.0475
1024	$9.7656E-04$	$9.65E-06$	2.0115	$2.37E-06$	2.0117
2048	$4.8828E-04$	$2.41E-06$	2.0029	$5.90E-07$	2.0029

Table 7.39: k-convergence for $\sigma_D = 0$, $\gamma = \gamma_1$, $Q = 5$, $d = 2$, $r = 1$. Errors e_{comb}

N	k	e_{comb}	EOC	residual
1	$1.0000E+00$	$8.60E+01$	-	8.3183e-14
2	$5.0000E-01$	$2.77E+01$	1.6349	2.4345e-14
4	$2.5000E-01$	$1.76E+01$	0.6506	4.9125e-14
8	$1.2500E-01$	$1.13E+01$	0.6454	5.6550e-14
16	$6.2500E-02$	$5.96E+00$	0.9205	9.4166e-14
32	$3.1250E-02$	$2.40E+00$	1.3106	9.4541e-14
64	$1.5625E-02$	$3.34E-02$	6.1678	9.3950e-14
128	$7.8125E-03$	$5.56E-03$	2.5885	9.3508e-14
256	$3.9062E-03$	$1.29E-03$	2.1024	9.9539e-14
512	$1.9531E-03$	$3.18E-04$	2.0237	1.0587e-13
1024	$9.7656E-04$	$7.92E-05$	2.0058	2.3518e-13
2048	$4.8828E-04$	$1.98E-05$	2.0014	4.8170e-13

Table 7.40: k-convergence for $\sigma_D = 0$, $\gamma = \gamma_1$, $Q = 5$, $d = 2$, $r = 1$. Errors e_0 and e_1

N	k	e_0	EOC	e_1	EOC
1	$1.0000E+00$	$6.54E+01$	-	$2.07E+01$	-
2	$5.0000E-01$	$2.29E+01$	1.5135	$4.82E+00$	2.1034
4	$2.5000E-01$	$1.45E+01$	0.6553	$3.11E+00$	0.6287
8	$1.2500E-01$	$9.20E+00$	0.6596	$2.08E+00$	0.5810
16	$6.2500E-02$	$4.80E+00$	0.9395	$1.16E+00$	0.8396
32	$3.1250E-02$	$1.91E+00$	1.3275	$4.92E-01$	1.2428
64	$1.5625E-02$	$2.68E-02$	6.1579	$6.65E-03$	6.2069
128	$7.8125E-03$	$4.48E-03$	2.5802	$1.08E-03$	2.6224
256	$3.9062E-03$	$1.04E-03$	2.1003	$2.50E-04$	2.1113
512	$1.9531E-03$	$2.57E-04$	2.0232	$6.14E-05$	2.0260
1024	$9.7656E-04$	$6.40E-05$	2.0057	$1.53E-05$	2.0064
2048	$4.8828E-04$	$1.60E-05$	2.0014	$3.82E-06$	2.0016

Table 7.41: k-convergence for $\sigma_D = 1$, $\gamma = \gamma_3$, $Q = 1$, $d = 1$, $r = 1$. Errors e_{comb}

N	k	e_{comb}	EOC	residual
1	$1.0000E+00$	$8.11E-01$	-	9.1551e-16
2	$5.0000E-01$	$6.69E-01$	0.2781	2.2377e-16
4	$2.5000E-01$	$3.69E-01$	0.8575	4.4755e-16
8	$1.2500E-01$	$1.47E-01$	1.3277	2.7014e-15
16	$6.2500E-02$	$4.66E-02$	1.6575	1.2562e-15
32	$3.1250E-02$	$1.31E-02$	1.8264	1.8310e-15
64	$1.5625E-02$	$3.40E-03$	1.9506	3.9721e-15
128	$7.8125E-03$	$8.50E-04$	2.0002	7.9441e-15
256	$3.9062E-03$	$2.11E-04$	2.0101	1.5888e-14
512	$1.9531E-03$	$5.24E-05$	2.0082	5.7286e-14
1024	$9.7656E-04$	$1.31E-05$	2.0050	1.1391e-13
2048	$4.8828E-04$	$3.26E-06$	2.0027	2.2782e-13

Table 7.42: k-convergence for $\sigma_D = 1$, $\gamma = \gamma_3$, $Q = 1$, $d = 1$, $r = 1$. Errors e_0 and e_1

N	k	e_0	EOC	e_1	EOC
1	$1.0000E+00$	$6.74E-01$	-	$1.36E-01$	-
2	$5.0000E-01$	$5.39E-01$	0.3225	$1.29E-01$	0.0764
4	$2.5000E-01$	$2.96E-01$	0.8641	$7.27E-02$	0.8302
8	$1.2500E-01$	$1.19E-01$	1.3118	$2.76E-02$	1.3945
16	$6.2500E-02$	$3.84E-02$	1.6344	$8.15E-03$	1.7616
32	$3.1250E-02$	$1.10E-02$	1.8111	$2.18E-03$	1.9008
64	$1.5625E-02$	$2.85E-03$	1.9437	$5.51E-04$	1.9855
128	$7.8125E-03$	$7.13E-04$	1.9978	$1.37E-04$	2.0123
256	$3.9062E-03$	$1.77E-04$	2.0094	$3.38E-05$	2.0138
512	$1.9531E-03$	$4.40E-05$	2.0080	$8.41E-06$	2.0093
1024	$9.7656E-04$	$1.10E-05$	2.0049	$2.09E-06$	2.0053
2048	$4.8828E-04$	$2.74E-06$	2.0027	$5.22E-07$	2.0029

Table 7.43: k-convergence for $\sigma_D = 1$, $\gamma = \gamma_3$, $Q = 1$, $d = 3$, $r = 1$. Errors e_{comb}

N	k	e_{comb}	EOC	residual
1	$1.0000E+00$	$2.28E-01$	-	3.3825e-15
2	$5.0000E-01$	$2.13E-01$	0.0999	9.2948e-15
4	$2.5000E-01$	$9.89E-02$	1.1086	2.2282e-14
8	$1.2500E-01$	$2.67E-02$	1.8876	2.1558e-14
16	$6.2500E-02$	$1.16E-02$	1.1981	1.5106e-14
32	$3.1250E-02$	$4.88E-03$	1.2553	1.6300e-14
64	$1.5625E-02$	$1.68E-03$	1.5346	2.9565e-14
128	$7.8125E-03$	$4.80E-04$	1.8114	3.3042e-14
256	$3.9062E-03$	$1.24E-04$	1.9497	8.6002e-14
512	$1.9531E-03$	$3.12E-05$	1.9928	1.8488e-13
1024	$9.7656E-04$	$7.80E-06$	2.0008	3.5726e-13
2048	$4.8828E-04$	$1.95E-06$	1.9970	9.1921e-13

Table 7.44: k-convergence for $\sigma_D = 1$, $\gamma = \gamma_3$, $Q = 1$, $d = 3$, $r = 1$. Errors e_0 and e_1

N	k	e_0	EOC	e_1	EOC
1	$1.0000E+00$	$1.95E-01$	-	$1.92E-02$	-
2	$5.0000E-01$	$1.67E-01$	0.2225	$2.65E-02$	-0.4636
4	$2.5000E-01$	$7.49E-02$	1.1592	$1.38E-02$	0.9379
8	$1.2500E-01$	$1.99E-02$	1.9124	$3.94E-03$	1.8125
16	$6.2500E-02$	$8.50E-03$	1.2263	$1.81E-03$	1.1187
32	$3.1250E-02$	$3.51E-03$	1.2754	$7.88E-04$	1.2023
64	$1.5625E-02$	$1.20E-03$	1.5445	$2.77E-04$	1.5094
128	$7.8125E-03$	$3.42E-04$	1.8152	$7.94E-05$	1.8020
256	$3.9062E-03$	$8.85E-05$	1.9508	$2.06E-05$	1.9467
512	$1.9531E-03$	$2.22E-05$	1.9932	$5.18E-06$	1.9919
1024	$9.7656E-04$	$5.56E-06$	2.0009	$1.29E-06$	2.0005
2048	$4.8828E-04$	$1.39E-06$	1.9971	$3.24E-07$	1.9969

Table 7.45: h-convergence for $\sigma_D = 1$, $\gamma = \gamma_3$, $Q = 1$, $d = 2$, $r = 1$, using $k = 1.0$. Errors e_{comb}

M	ndof	h	e_{comb}	EOC	residual
1	5	$1.0000E+00$	$1.05E+00$	-	2.4460e-15
2	13	$5.0000E-01$	$4.99E-01$	1.0711	4.7779e-16
4	41	$2.5000E-01$	$1.42E-01$	1.8158	9.7039e-16
8	145	$1.2500E-01$	$3.52E-02$	2.0077	1.2891e-15
16	545	$6.2500E-02$	$8.88E-03$	1.9872	2.5844e-15
32	2113	$3.1250E-02$	$2.23E-03$	1.9964	4.9291e-15
64	8321	$1.5625E-02$	$5.57E-04$	1.9987	3.9935e-14
128	33025	$7.8125E-03$	$1.39E-04$	1.9994	6.9165e-10
256	131585	$3.9062E-03$	$3.59E-05$	1.9568	1.7508e-08

Table 7.46: h-convergence for $\sigma_D = 1$, $\gamma = \gamma_3$, $Q = 1$, $d = 2$, $r = 1$, using $k = 1.0$. Errors e_0 and e_1

M	ndof	h	e_0	EOC	e_1	EOC
1	5	$1.0000E+00$	$1.38E-01$	-	$9.10E-01$	-
2	13	$5.0000E-01$	$1.09E-01$	0.3395	$7.79E-01$	0.2236
4	41	$2.5000E-01$	$3.35E-02$	1.7059	$4.33E-01$	0.8482
8	145	$1.2500E-01$	$7.54E-03$	2.1501	$2.21E-01$	0.9663
16	545	$6.2500E-02$	$1.92E-03$	1.9754	$1.11E-01$	0.9904
32	2113	$3.1250E-02$	$4.81E-04$	1.9938	$5.59E-02$	0.9972
64	8321	$1.5625E-02$	$1.21E-04$	1.9972	$2.79E-02$	0.9992
128	33025	$7.8125E-03$	$3.02E-05$	1.9983	$1.40E-02$	0.9998
256	131585	$3.9062E-03$	$8.60E-06$	1.8106	$6.99E-03$	0.9999

Table 7.47: h-convergence for $\sigma_D = 1$, $\gamma = \gamma_3$, $Q = 2$, $d = 2$, $r = 1$, using $k = 1.0$. Errors e_{comb}

M	ndof	h	e_{comb}	EOC	residual
1	5	$1.0000E+00$	$1.05E+00$	-	2.4460e-15
2	13	$5.0000E-01$	$4.99E-01$	1.0711	4.7779e-16
4	41	$2.5000E-01$	$1.42E-01$	1.8158	9.7039e-16
8	145	$1.2500E-01$	$3.52E-02$	2.0077	1.2891e-15
16	545	$6.2500E-02$	$8.88E-03$	1.9872	2.5844e-15
32	2113	$3.1250E-02$	$2.23E-03$	1.9964	4.9291e-15
64	8321	$1.5625E-02$	$5.57E-04$	1.9987	3.9935e-14
128	33025	$7.8125E-03$	$1.39E-04$	1.9994	6.9165e-10
256	131585	$3.9062E-03$	$3.59E-05$	1.9568	1.7508e-08

Table 7.48: h-convergence for $\sigma_D = 1$, $\gamma = \gamma_3$, $Q = 2$, $d = 2$, $r = 1$, using $k = 1.0$. Errors e_0 and e_1

M	ndof	h	e_0	EOC	e_1	EOC
1	5	$1.0000E+00$	$1.38E-01$	-	$9.10E-01$	-
2	13	$5.0000E-01$	$1.09E-01$	0.3395	$7.79E-01$	0.2236
4	41	$2.5000E-01$	$3.35E-02$	1.7059	$4.33E-01$	0.8482
8	145	$1.2500E-01$	$7.54E-03$	2.1501	$2.21E-01$	0.9663
16	545	$6.2500E-02$	$1.92E-03$	1.9754	$1.11E-01$	0.9904
32	2113	$3.1250E-02$	$4.81E-04$	1.9938	$5.59E-02$	0.9972
64	8321	$1.5625E-02$	$1.21E-04$	1.9972	$2.79E-02$	0.9992
128	33025	$7.8125E-03$	$3.02E-05$	1.9983	$1.40E-02$	0.9998
256	131585	$3.9062E-03$	$8.60E-06$	1.8106	$6.99E-03$	0.9999

Table 7.49: h-convergence for $\sigma_D = 1$, $\gamma = \gamma_3$, $Q = 3$, $d = 2$, $r = 1$, using $k = 1.0$. Errors e_{comb}

M	ndof	h	e_{comb}	EOC	residual
1	5	$1.0000E+00$	$1.24E+00$	-	1.7936e-15
2	13	$5.0000E-01$	$6.41E-01$	0.9464	1.3755e-15
4	41	$2.5000E-01$	$2.07E-01$	1.6315	2.9785e-15
8	145	$1.2500E-01$	$4.27E-02$	2.2787	9.1954e-15
16	545	$6.2500E-02$	$1.26E-02$	1.7627	2.1446e-14
32	2113	$3.1250E-02$	$3.26E-03$	1.9469	5.9457e-14
64	8321	$1.5625E-02$	$8.29E-04$	1.9754	1.6167e-13
128	33025	$7.8125E-03$	$2.08E-04$	1.9936	8.4979e-13
256	131585	$3.9062E-03$	$5.21E-05$	1.9982	1.3393e-09

Table 7.50: h-convergence for $\sigma_D = 1$, $\gamma = \gamma_3$, $Q = 3$, $d = 2$, $r = 1$, using $k = 1.0$. Errors e_0 and e_1

M	ndof	h	e_0	EOC	e_1	EOC
1	5	$1.0000E+00$	$3.26E-01$	-	$9.10E-01$	-
2	13	$5.0000E-01$	$2.52E-01$	0.3720	$7.79E-01$	0.2239
4	41	$2.5000E-01$	$9.88E-02$	1.3488	$4.33E-01$	0.8487
8	145	$1.2500E-01$	$1.50E-02$	2.7221	$2.21E-01$	0.9664
16	545	$6.2500E-02$	$5.60E-03$	1.4180	$1.11E-01$	0.9904
32	2113	$3.1250E-02$	$1.52E-03$	1.8868	$5.59E-02$	0.9972
64	8321	$1.5625E-02$	$3.93E-04$	1.9486	$2.79E-02$	0.9992
128	33025	$7.8125E-03$	$9.90E-05$	1.9868	$1.40E-02$	0.9998
256	131585	$3.9062E-03$	$2.48E-05$	1.9964	$6.99E-03$	0.9999

Table 7.51: h-convergence for $\sigma_D = 1$, $\gamma = \gamma_3$, $Q = 4$, $d = 2$, $r = 1$, using $k = 0.0001$. Errors e_{comb}

M	ndof	h	e_{comb}	EOC	residual
1	5	$1.0000E+00$	$1.16E+00$	-	7.5557e-12
2	13	$5.0000E-01$	$5.51E-01$	1.0743	3.7941e-12
4	41	$2.5000E-01$	$1.55E-01$	1.8323	1.2966e-12
8	145	$1.2500E-01$	$8.49E-02$	0.8644	6.2113e-13
16	545	$6.2500E-02$	$5.62E-02$	0.5968	2.1823e-13

Table 7.52: h-convergence for $\sigma_D = 1$, $\gamma = \gamma_3$, $Q = 4$, $d = 2$, $r = 1$, using $k = 0.0001$. Errors e_0 and e_1

M	ndof	h	e_0	EOC	e_1	EOC
1	5	$1.0000E+00$	$1.90E-01$	-	$9.69E-01$	-
2	13	$5.0000E-01$	$1.38E-01$	0.4648	$8.26E-01$	0.2313
4	41	$2.5000E-01$	$4.38E-02$	1.6530	$4.43E-01$	0.8974
8	145	$1.2500E-01$	$5.67E-02$	-0.3726	$2.26E-01$	0.9727
16	545	$6.2500E-02$	$4.91E-02$	0.2087	$1.14E-01$	0.9913

Table 7.53: h-convergence for $\sigma_D = 1$, $\gamma = \gamma_3$, $Q = 5$, $d = 2$, $r = 1$, using $k = 0.0001$. Errors e_{comb}

M	ndof	h	e_{comb}	EOC	residual
1	5	$1.0000E+00$	$1.16E+00$	-	7.4030e-12
2	13	$5.0000E-01$	$5.51E-01$	1.0743	3.9189e-12
4	41	$2.5000E-01$	$1.55E-01$	1.8323	1.0961e-12
8	145	$1.2500E-01$	$8.49E-02$	0.8644	6.1319e-13
16	545	$6.2500E-02$	$5.62E-02$	0.5968	2.2186e-13

Table 7.54: h-convergence for $\sigma_D = 1$, $\gamma = \gamma_3$, $Q = 5$, $d = 2$, $r = 1$, using $k = 0.0001$. Errors e_0 and e_1

M	ndof	h	e_0	EOC	e_1	EOC
1	5	$1.0000E+00$	$1.90E-01$	-	$9.69E-01$	-
2	13	$5.0000E-01$	$1.38E-01$	0.4648	$8.26E-01$	0.2313
4	41	$2.5000E-01$	$4.38E-02$	1.6530	$4.43E-01$	0.8974
8	145	$1.2500E-01$	$5.67E-02$	-0.3726	$2.26E-01$	0.9727
16	545	$6.2500E-02$	$4.91E-02$	0.2087	$1.14E-01$	0.9913

Table 7.55: h-convergence for $\sigma_D = 1$, $\gamma = \gamma_1$, $Q = 1$, $d = 2$, $r = 1$, using $k = 1.0$. Errors e_{comb}

M	ndof	h	e_{comb}	EOC	residual
1	5	$1.0000E+00$	$1.04E+00$	-	2.2430e-15
2	13	$5.0000E-01$	$4.73E-01$	1.1334	3.5547e-16
4	41	$2.5000E-01$	$1.36E-01$	1.7962	5.5870e-16
8	145	$1.2500E-01$	$3.52E-02$	1.9531	1.2688e-15
16	545	$6.2500E-02$	$8.88E-03$	1.9869	2.5544e-15
32	2113	$3.1250E-02$	$2.23E-03$	1.9960	5.0299e-15
64	8321	$1.5625E-02$	$5.57E-04$	1.9987	1.6137e-14
128	33025	$7.8125E-03$	$1.39E-04$	1.9983	6.9700e-09
256	131585	$3.9062E-03$	$3.56E-05$	1.9683	1.3676e-08

Table 7.56: h-convergence for $\sigma_D = 1$, $\gamma = \gamma_1$, $Q = 1$, $d = 2$, $r = 1$, using $k = 1.0$. Errors e_0 and e_1

M	ndof	h	e_0	EOC	e_1	EOC
1	5	$1.0000E+00$	$1.29E-01$	-	$9.10E-01$	-
2	13	$5.0000E-01$	$8.38E-02$	0.6185	$7.79E-01$	0.2236
4	41	$2.5000E-01$	$2.81E-02$	1.5763	$4.33E-01$	0.8482
8	145	$1.2500E-01$	$7.51E-03$	1.9034	$2.21E-01$	0.9663
16	545	$6.2500E-02$	$1.91E-03$	1.9740	$1.11E-01$	0.9904
32	2113	$3.1250E-02$	$4.81E-04$	1.9916	$5.59E-02$	0.9972
64	8321	$1.5625E-02$	$1.20E-04$	1.9970	$2.79E-02$	0.9992
128	33025	$7.8125E-03$	$3.03E-05$	1.9929	$1.40E-02$	0.9998
256	131585	$3.9062E-03$	$8.34E-06$	1.8596	$6.99E-03$	0.9999

Table 7.57: h-convergence for $\sigma_D = 1$, $\gamma = \gamma_1$, $Q = 2$, $d = 2$, $r = 1$, using $k = 1.0$. Errors e_{comb}

M	ndof	h	e_{comb}	EOC	residual
1	5	$1.0000E+00$	$1.04E+00$	-	2.2430e-15
2	13	$5.0000E-01$	$4.73E-01$	1.1334	3.5547e-16
4	41	$2.5000E-01$	$1.36E-01$	1.7962	5.5870e-16
8	145	$1.2500E-01$	$3.52E-02$	1.9531	1.2688e-15
16	545	$6.2500E-02$	$8.88E-03$	1.9869	2.5544e-15
32	2113	$3.1250E-02$	$2.23E-03$	1.9960	5.0299e-15
64	8321	$1.5625E-02$	$5.57E-04$	1.9987	1.6137e-14
128	33025	$7.8125E-03$	$1.39E-04$	1.9983	6.9700e-09
256	131585	$3.9062E-03$	$3.56E-05$	1.9683	1.3676e-08

Table 7.58: h-convergence for $\sigma_D = 1$, $\gamma = \gamma_1$, $Q = 2$, $d = 2$, $r = 1$, using $k = 1.0$. Errors e_0 and e_1

M	ndof	h	e_0	EOC	e_1	EOC
1	5	$1.0000E+00$	$1.29E-01$	-	$9.10E-01$	-
2	13	$5.0000E-01$	$8.38E-02$	0.6185	$7.79E-01$	0.2236
4	41	$2.5000E-01$	$2.81E-02$	1.5763	$4.33E-01$	0.8482
8	145	$1.2500E-01$	$7.51E-03$	1.9034	$2.21E-01$	0.9663
16	545	$6.2500E-02$	$1.91E-03$	1.9740	$1.11E-01$	0.9904
32	2113	$3.1250E-02$	$4.81E-04$	1.9916	$5.59E-02$	0.9972
64	8321	$1.5625E-02$	$1.20E-04$	1.9970	$2.79E-02$	0.9992
128	33025	$7.8125E-03$	$3.03E-05$	1.9929	$1.40E-02$	0.9998
256	131585	$3.9062E-03$	$8.34E-06$	1.8596	$6.99E-03$	0.9999

Table 7.59: h-convergence for $\sigma_D = 1$, $\gamma = \gamma_1$, $Q = 3$, $d = 2$, $r = 1$, using $k = 1.0$. Errors e_{comb}

M	ndof	h	e_{comb}	EOC	residual
1	5	$1.0000E+00$	$1.15E+00$	-	5.6610e-16
2	13	$5.0000E-01$	$5.54E-01$	1.0525	1.9445e-15
4	41	$2.5000E-01$	$1.44E-01$	1.9431	3.7352e-15
8	145	$1.2500E-01$	$3.84E-02$	1.9067	1.1407e-14
16	545	$6.2500E-02$	$9.70E-03$	1.9847	2.2694e-14
32	2113	$3.1250E-02$	$2.43E-03$	1.9946	6.4070e-14
64	8321	$1.5625E-02$	$6.09E-04$	1.9984	1.6559e-13
128	33025	$7.8125E-03$	$1.52E-04$	1.9995	2.2813e-11
256	131585	$3.9062E-03$	$3.81E-05$	1.9999	6.1407e-09

Table 7.60: h-convergence for $\sigma_D = 1$, $\gamma = \gamma_1$, $Q = 3$, $d = 2$, $r = 1$, using $k = 1.0$. Errors e_0 and e_1

M	ndof	h	e_0	EOC	e_1	EOC
1	5	$1.0000E+00$	$2.38E-01$	-	$9.10E-01$	-
2	13	$5.0000E-01$	$1.64E-01$	0.5399	$7.79E-01$	0.2236
4	41	$2.5000E-01$	$3.58E-02$	2.1969	$4.33E-01$	0.8483
8	145	$1.2500E-01$	$1.07E-02$	1.7402	$2.21E-01$	0.9663
16	545	$6.2500E-02$	$2.73E-03$	1.9701	$1.11E-01$	0.9904
32	2113	$3.1250E-02$	$6.89E-04$	1.9881	$5.59E-02$	0.9972
64	8321	$1.5625E-02$	$1.73E-04$	1.9965	$2.79E-02$	0.9992
128	33025	$7.8125E-03$	$4.32E-05$	1.9989	$1.40E-02$	0.9998
256	131585	$3.9062E-03$	$1.08E-05$	1.9999	$6.99E-03$	0.9999

Table 7.61: h-convergence for $\sigma_D = 1$, $\gamma = \gamma_1$, $Q = 4$, $d = 2$, $r = 1$, using $k = 0.0001$. Errors e_{comb}

M	ndof	h	e_{comb}	EOC	residual
1	5	$1.0000E+00$	$1.15E+00$	-	7.4038e-12
2	13	$5.0000E-01$	$4.96E-01$	1.2136	3.6672e-12
4	41	$2.5000E-01$	$1.46E-01$	1.7683	2.1491e-12
8	145	$1.2500E-01$	$3.92E-02$	1.8938	5.8784e-13
16	545	$6.2500E-02$	$9.88E-03$	1.9866	4.0683e-13

Table 7.62: h-convergence for $\sigma_D = 1$, $\gamma = \gamma_1$, $Q = 4$, $d = 2$, $r = 1$, using $k = 0.0001$. Errors e_0 and e_1

M	ndof	h	e_0	EOC	e_1	EOC
1	5	$1.0000E+00$	$1.80E-01$	-	$9.69E-01$	-
2	13	$5.0000E-01$	$8.28E-02$	1.1221	$8.26E-01$	0.2313
4	41	$2.5000E-01$	$3.47E-02$	1.2559	$4.43E-01$	0.8973
8	145	$1.2500E-01$	$1.09E-02$	1.6665	$2.26E-01$	0.9730
16	545	$6.2500E-02$	$2.78E-03$	1.9731	$1.14E-01$	0.9918

Table 7.63: h-convergence for $\sigma_D = 1$, $\gamma = \gamma_1$, $Q = 5$, $d = 2$, $r = 1$, using $k = 0.0001$. Errors e_{comb}

M	ndof	h	e_{comb}	EOC	residual
1	5	$1.0000E+00$	$1.15E+00$	-	7.5559e-12
2	13	$5.0000E-01$	$4.96E-01$	1.2136	3.1588e-12
4	41	$2.5000E-01$	$1.46E-01$	1.7683	1.9952e-12
8	145	$1.2500E-01$	$3.92E-02$	1.8938	5.5730e-13
16	545	$6.2500E-02$	$9.88E-03$	1.9866	4.1491e-13

Table 7.64: h-convergence for $\sigma_D = 1$, $\gamma = \gamma_1$, $Q = 5$, $d = 2$, $r = 1$, using $k = 0.0001$. Errors e_0 and e_1

M	ndof	h	e_0	EOC	e_1	EOC
1	5	$1.0000E+00$	$1.80E-01$	-	$9.69E-01$	-
2	13	$5.0000E-01$	$8.28E-02$	1.1221	$8.26E-01$	0.2313
4	41	$2.5000E-01$	$3.47E-02$	1.2559	$4.43E-01$	0.8973
8	145	$1.2500E-01$	$1.09E-02$	1.6665	$2.26E-01$	0.9730
16	545	$6.2500E-02$	$2.78E-03$	1.9731	$1.14E-01$	0.9918

Table 7.65: h-convergence for $\sigma_D = 0$, $\gamma = \gamma_3$, $Q = 1$, $d = 2$, $r = 1$, using $k = 1.0$. Errors e_{comb}

M	ndof	h	e_{comb}	EOC	residual
1	5	$1.0000E+00$	$1.04E+00$	-	1.5685e-15
2	13	$5.0000E-01$	$4.98E-01$	1.0583	3.5141e-16
4	41	$2.5000E-01$	$1.42E-01$	1.8109	9.4350e-16
8	145	$1.2500E-01$	$3.53E-02$	2.0098	1.2364e-15
16	545	$6.2500E-02$	$8.89E-03$	1.9880	2.6103e-15
32	2113	$3.1250E-02$	$2.23E-03$	1.9967	4.8869e-15
64	8321	$1.5625E-02$	$5.57E-04$	1.9989	3.9330e-14
128	33025	$7.8125E-03$	$1.39E-04$	1.9996	1.5702e-10
256	131585	$3.9062E-03$	$3.49E-05$	1.9964	2.1122e-09

Table 7.66: h-convergence for $\sigma_D = 0$, $\gamma = \gamma_3$, $Q = 1$, $d = 2$, $r = 1$, using $k = 1.0$. Errors e_0 and e_1

M	ndof	h	e_0	EOC	e_1	EOC
1	5	$1.0000E+00$	$1.28E-01$	-	$9.10E-01$	-
2	13	$5.0000E-01$	$1.09E-01$	0.2351	$7.79E-01$	0.2233
4	41	$2.5000E-01$	$3.36E-02$	1.6911	$4.33E-01$	0.8462
8	145	$1.2500E-01$	$7.56E-03$	2.1538	$2.22E-01$	0.9678
16	545	$6.2500E-02$	$1.92E-03$	1.9776	$1.11E-01$	0.9908
32	2113	$3.1250E-02$	$4.81E-04$	1.9948	$5.59E-02$	0.9972
64	8321	$1.5625E-02$	$1.21E-04$	1.9977	$2.79E-02$	0.9992
128	33025	$7.8125E-03$	$3.02E-05$	1.9990	$1.40E-02$	0.9998
256	131585	$3.9062E-03$	$7.63E-06$	1.9839	$6.99E-03$	0.9999

Table 7.67: h-convergence for $\sigma_D = 0$, $\gamma = \gamma_3$, $Q = 2$, $d = 2$, $r = 1$, using $k = 1.0$. Errors e_{comb}

M	ndof	h	e_{comb}	EOC	residual
1	5	$1.0000E+00$	$1.04E+00$	-	1.5685e-15
2	13	$5.0000E-01$	$4.98E-01$	1.0583	3.5141e-16
4	41	$2.5000E-01$	$1.42E-01$	1.8109	9.4350e-16
8	145	$1.2500E-01$	$3.53E-02$	2.0098	1.2364e-15
16	545	$6.2500E-02$	$8.89E-03$	1.9880	2.6103e-15
32	2113	$3.1250E-02$	$2.23E-03$	1.9967	4.8869e-15
64	8321	$1.5625E-02$	$5.57E-04$	1.9989	3.9330e-14
128	33025	$7.8125E-03$	$1.39E-04$	1.9996	1.5702e-10
256	131585	$3.9062E-03$	$3.49E-05$	1.9964	2.1122e-09

Table 7.68: h-convergence for $\sigma_D = 0$, $\gamma = \gamma_3$, $Q = 2$, $d = 2$, $r = 1$, using $k = 1.0$. Errors e_0 and e_1

M	ndof	h	e_0	EOC	e_1	EOC
1	5	$1.0000E+00$	$1.28E-01$	-	$9.10E-01$	-
2	13	$5.0000E-01$	$1.09E-01$	0.2351	$7.79E-01$	0.2233
4	41	$2.5000E-01$	$3.36E-02$	1.6911	$4.33E-01$	0.8462
8	145	$1.2500E-01$	$7.56E-03$	2.1538	$2.22E-01$	0.9678
16	545	$6.2500E-02$	$1.92E-03$	1.9776	$1.11E-01$	0.9908
32	2113	$3.1250E-02$	$4.81E-04$	1.9948	$5.59E-02$	0.9972
64	8321	$1.5625E-02$	$1.21E-04$	1.9977	$2.79E-02$	0.9992
128	33025	$7.8125E-03$	$3.02E-05$	1.9990	$1.40E-02$	0.9998
256	131585	$3.9062E-03$	$7.63E-06$	1.9839	$6.99E-03$	0.9999

Table 7.69: h-convergence for $\sigma_D = 0$, $\gamma = \gamma_3$, $Q = 3$, $d = 2$, $r = 1$, using $k = 1.0$. Errors e_{comb}

M	ndof	h	e_{comb}	EOC	residual
1	5	$1.0000E+00$	$1.24E+00$	-	4.9651e-16
2	13	$5.0000E-01$	$6.50E-01$	0.9275	1.6222e-15
4	41	$2.5000E-01$	$2.07E-01$	1.6497	3.4993e-15
8	145	$1.2500E-01$	$4.27E-02$	2.2798	8.6033e-15
16	545	$6.2500E-02$	$1.26E-02$	1.7633	2.1477e-14
32	2113	$3.1250E-02$	$3.26E-03$	1.9470	6.1494e-14
64	8321	$1.5625E-02$	$8.29E-04$	1.9755	1.5788e-13
128	33025	$7.8125E-03$	$2.08E-04$	1.9936	8.5481e-13
256	131585	$3.9062E-03$	$5.21E-05$	1.9983	1.3393e-09

Table 7.70: h-convergence for $\sigma_D = 0$, $\gamma = \gamma_3$, $Q = 3$, $d = 2$, $r = 1$, using $k = 1.0$. Errors e_0 and e_1

M	ndof	h	e_0	EOC	e_1	EOC
1	5	$1.0000E+00$	$3.27E-01$	-	$9.10E-01$	-
2	13	$5.0000E-01$	$2.54E-01$	0.3630	$7.92E-01$	0.1998
4	41	$2.5000E-01$	$9.88E-02$	1.3628	$4.34E-01$	0.8692
8	145	$1.2500E-01$	$1.50E-02$	2.7214	$2.22E-01$	0.9692
16	545	$6.2500E-02$	$5.60E-03$	1.4188	$1.11E-01$	0.9908
32	2113	$3.1250E-02$	$1.52E-03$	1.8870	$5.59E-02$	0.9972
64	8321	$1.5625E-02$	$3.93E-04$	1.9487	$2.79E-02$	0.9992
128	33025	$7.8125E-03$	$9.90E-05$	1.9868	$1.40E-02$	0.9998
256	131585	$3.9062E-03$	$2.48E-05$	1.9964	$6.99E-03$	0.9999

Table 7.71: h-convergence for $\sigma_D = 0$, $\gamma = \gamma_3$, $Q = 4$, $d = 2$, $r = 1$, using $k = 0.0001$. Errors e_{comb}

M	ndof	h	e_{comb}	EOC	residual
1	5	$1.0000E+00$	$1.12E+00$	-	7.5035e-12
2	13	$5.0000E-01$	$5.36E-01$	1.0680	3.7924e-12
4	41	$2.5000E-01$	$1.60E-01$	1.7434	1.3101e-12
8	145	$1.2500E-01$	$8.54E-02$	0.9078	6.0765e-13
16	545	$6.2500E-02$	$5.67E-02$	0.5906	2.2682e-13

Table 7.72: h-convergence for $\sigma_D = 0$, $\gamma = \gamma_3$, $Q = 4$, $d = 2$, $r = 1$, using $k = 0.0001$. Errors e_0 and e_1

M	ndof	h	e_0	EOC	e_1	EOC
1	5	$1.0000E+00$	$1.62E-01$	-	$9.63E-01$	-
2	13	$5.0000E-01$	$1.20E-01$	0.4323	$8.33E-01$	0.2087
4	41	$2.5000E-01$	$4.88E-02$	1.2959	$4.46E-01$	0.9027
8	145	$1.2500E-01$	$5.71E-02$	-0.2278	$2.26E-01$	0.9790
16	545	$6.2500E-02$	$4.96E-02$	0.2038	$1.14E-01$	0.9923

Table 7.73: h-convergence for $\sigma_D = 0$, $\gamma = \gamma_3$, $Q = 5$, $d = 2$, $r = 1$, using $k = 0.0001$. Errors e_{comb}

M	ndof	h	e_{comb}	EOC	residual
1	5	$1.0000E+00$	$1.12E+00$	-	5.9642e-12
2	13	$5.0000E-01$	$5.36E-01$	1.0680	3.4073e-12
4	41	$2.5000E-01$	$1.60E-01$	1.7434	1.3662e-12
8	145	$1.2500E-01$	$8.54E-02$	0.9078	6.2500e-13
16	545	$6.2500E-02$	$5.67E-02$	0.5906	2.1125e-13

Table 7.74: h-convergence for $\sigma_D = 0$, $\gamma = \gamma_3$, $Q = 5$, $d = 2$, $r = 1$, using $k = 0.0001$. Errors e_0 and e_1

M	ndof	h	e_0	EOC	e_1	EOC
1	5	$1.0000E+00$	$1.62E-01$	-	$9.63E-01$	-
2	13	$5.0000E-01$	$1.20E-01$	0.4323	$8.33E-01$	0.2087
4	41	$2.5000E-01$	$4.88E-02$	1.2959	$4.46E-01$	0.9027
8	145	$1.2500E-01$	$5.71E-02$	-0.2278	$2.26E-01$	0.9790
16	545	$6.2500E-02$	$4.96E-02$	0.2038	$1.14E-01$	0.9923

Table 7.75: h-convergence for $\sigma_D = 0$, $\gamma = \gamma_1$, $Q = 1$, $d = 2$, $r = 1$, using $k = 1.0$. Errors e_{comb}

M	ndof	h	e_{comb}	EOC	residual
1	5	$1.0000E+00$	$1.03E+00$	-	4.7247e-15
2	13	$5.0000E-01$	$4.76E-01$	1.1179	1.9290e-16
4	41	$2.5000E-01$	$1.37E-01$	1.8010	7.2706e-16
8	145	$1.2500E-01$	$3.52E-02$	1.9561	1.3316e-15
16	545	$6.2500E-02$	$8.88E-03$	1.9877	2.4187e-15
32	2113	$3.1250E-02$	$2.23E-03$	1.9962	4.9266e-15
64	8321	$1.5625E-02$	$5.57E-04$	1.9988	1.8490e-14
128	33025	$7.8125E-03$	$1.39E-04$	1.9996	2.5348e-10
256	131585	$3.9062E-03$	$3.49E-05$	1.9975	1.5260e-09

Table 7.76: h-convergence for $\sigma_D = 0$, $\gamma = \gamma_1$, $Q = 1$, $d = 2$, $r = 1$, using $k = 1.0$. Errors e_0 and e_1

M	ndof	h	e_0	EOC	e_1	EOC
1	5	$1.0000E+00$	$1.24E-01$	-	$9.10E-01$	-
2	13	$5.0000E-01$	$8.45E-02$	0.5541	$7.84E-01$	0.2152
4	41	$2.5000E-01$	$2.83E-02$	1.5771	$4.33E-01$	0.8542
8	145	$1.2500E-01$	$7.53E-03$	1.9107	$2.22E-01$	0.9681
16	545	$6.2500E-02$	$1.91E-03$	1.9762	$1.11E-01$	0.9908
32	2113	$3.1250E-02$	$4.81E-04$	1.9926	$5.59E-02$	0.9972
64	8321	$1.5625E-02$	$1.20E-04$	1.9975	$2.79E-02$	0.9992
128	33025	$7.8125E-03$	$3.01E-05$	1.9989	$1.40E-02$	0.9998
256	131585	$3.9062E-03$	$7.59E-06$	1.9889	$6.99E-03$	0.9999

Table 7.77: h-convergence for $\sigma_D = 0$, $\gamma = \gamma_1$, $Q = 2$, $d = 2$, $r = 1$, using $k = 1.0$. Errors e_{comb}

M	ndof	h	e_{comb}	EOC	residual
1	5	$1.0000E+00$	$1.03E+00$	-	4.7247e-15
2	13	$5.0000E-01$	$4.76E-01$	1.1179	1.9290e-16
4	41	$2.5000E-01$	$1.37E-01$	1.8010	7.2706e-16
8	145	$1.2500E-01$	$3.52E-02$	1.9561	1.3316e-15
16	545	$6.2500E-02$	$8.88E-03$	1.9877	2.4187e-15
32	2113	$3.1250E-02$	$2.23E-03$	1.9962	4.9266e-15
64	8321	$1.5625E-02$	$5.57E-04$	1.9988	1.8490e-14
128	33025	$7.8125E-03$	$1.39E-04$	1.9996	2.5348e-10
256	131585	$3.9062E-03$	$3.49E-05$	1.9975	1.5260e-09

Table 7.78: h-convergence for $\sigma_D = 0$, $\gamma = \gamma_1$, $Q = 2$, $d = 2$, $r = 1$, using $k = 1.0$. Errors e_0 and e_1

M	ndof	h	e_0	EOC	e_1	EOC
1	5	$1.0000E+00$	$1.24E-01$	-	$9.10E-01$	-
2	13	$5.0000E-01$	$8.45E-02$	0.5541	$7.84E-01$	0.2152
4	41	$2.5000E-01$	$2.83E-02$	1.5771	$4.33E-01$	0.8542
8	145	$1.2500E-01$	$7.53E-03$	1.9107	$2.22E-01$	0.9681
16	545	$6.2500E-02$	$1.91E-03$	1.9762	$1.11E-01$	0.9908
32	2113	$3.1250E-02$	$4.81E-04$	1.9926	$5.59E-02$	0.9972
64	8321	$1.5625E-02$	$1.20E-04$	1.9975	$2.79E-02$	0.9992
128	33025	$7.8125E-03$	$3.01E-05$	1.9989	$1.40E-02$	0.9998
256	131585	$3.9062E-03$	$7.59E-06$	1.9889	$6.99E-03$	0.9999

Table 7.79: h-convergence for $\sigma_D = 0$, $\gamma = \gamma_1$, $Q = 3$, $d = 2$, $r = 1$, using $k = 1.0$. Errors e_{comb}

M	ndof	h	e_{comb}	EOC	residual
1	5	$1.0000E+00$	$1.16E+00$	-	7.8505e-16
2	13	$5.0000E-01$	$5.69E-01$	1.0299	1.6373e-15
4	41	$2.5000E-01$	$1.44E-01$	1.9774	4.8420e-15
8	145	$1.2500E-01$	$3.84E-02$	1.9102	9.9108e-15
16	545	$6.2500E-02$	$9.70E-03$	1.9853	2.1942e-14
32	2113	$3.1250E-02$	$2.43E-03$	1.9948	6.4449e-14
64	8321	$1.5625E-02$	$6.09E-04$	1.9985	1.6858e-13
128	33025	$7.8125E-03$	$1.52E-04$	1.9996	2.2482e-11
256	131585	$3.9062E-03$	$3.81E-05$	2.0000	6.3607e-09

Table 7.80: h-convergence for $\sigma_D = 0$, $\gamma = \gamma_1$, $Q = 3$, $d = 2$, $r = 1$, using $k = 1.0$. Errors e_0 and e_1

M	ndof	h	e_0	EOC	e_1	EOC
1	5	$1.0000E+00$	$2.52E-01$	-	$9.09E-01$	-
2	13	$5.0000E-01$	$1.69E-01$	0.5774	$8.00E-01$	0.1852
4	41	$2.5000E-01$	$3.61E-02$	2.2273	$4.33E-01$	0.8836
8	145	$1.2500E-01$	$1.07E-02$	1.7490	$2.22E-01$	0.9681
16	545	$6.2500E-02$	$2.74E-03$	1.9713	$1.11E-01$	0.9908
32	2113	$3.1250E-02$	$6.89E-04$	1.9887	$5.59E-02$	0.9972
64	8321	$1.5625E-02$	$1.73E-04$	1.9968	$2.79E-02$	0.9992
128	33025	$7.8125E-03$	$4.32E-05$	1.9991	$1.40E-02$	0.9998
256	131585	$3.9062E-03$	$1.08E-05$	2.0000	$6.99E-03$	0.9999

Table 7.81: h-convergence for $\sigma_D = 0$, $\gamma = \gamma_1$, $Q = 4$, $d = 2$, $r = 1$, using $k = 0.0001$. Errors e_{comb}

M	ndof	h	e_{comb}	EOC	residual
1	5	$1.0000E+00$	$1.13E+00$	-	7.7215e-12
2	13	$5.0000E-01$	$5.07E-01$	1.1505	3.8616e-12
4	41	$2.5000E-01$	$1.52E-01$	1.7408	2.1067e-12
8	145	$1.2500E-01$	$4.09E-02$	1.8918	5.7619e-13
16	545	$6.2500E-02$	$1.03E-02$	1.9854	4.1770e-13

Table 7.82: h-convergence for $\sigma_D = 0$, $\gamma = \gamma_1$, $Q = 4$, $d = 2$, $r = 1$, using $k = 0.0001$. Errors e_0 and e_1

M	ndof	h	e_0	EOC	e_1	EOC
1	5	$1.0000E+00$	$1.63E-01$	-	$9.63E-01$	-
2	13	$5.0000E-01$	$8.08E-02$	1.0092	$8.53E-01$	0.1758
4	41	$2.5000E-01$	$4.05E-02$	0.9987	$4.45E-01$	0.9375
8	145	$1.2500E-01$	$1.26E-02$	1.6782	$2.26E-01$	0.9781
16	545	$6.2500E-02$	$3.23E-03$	1.9687	$1.14E-01$	0.9929

Table 7.83: h-convergence for $\sigma_D = 0$, $\gamma = \gamma_1$, $Q = 5$, $d = 2$, $r = 1$, using $k = 0.0001$. Errors e_{comb}

M	ndof	h	e_{comb}	EOC	residual
1	5	$1.0000E+00$	$1.13E+00$	-	7.6272e-12
2	13	$5.0000E-01$	$5.07E-01$	1.1505	3.8478e-12
4	41	$2.5000E-01$	$1.52E-01$	1.7408	2.1099e-12
8	145	$1.2500E-01$	$4.09E-02$	1.8918	5.9284e-13
16	545	$6.2500E-02$	$1.03E-02$	1.9854	3.9746e-13

Table 7.84: h-convergence for $\sigma_D = 0$, $\gamma = \gamma_1$, $Q = 5$, $d = 2$, $r = 1$, using $k = 0.0001$. Errors e_0 and e_1

M	ndof	h	e_0	EOC	e_1	EOC
1	5	$1.0000E+00$	$1.63E-01$	-	$9.63E-01$	-
2	13	$5.0000E-01$	$8.08E-02$	1.0092	$8.53E-01$	0.1758
4	41	$2.5000E-01$	$4.05E-02$	0.9987	$4.45E-01$	0.9375
8	145	$1.2500E-01$	$1.26E-02$	1.6782	$2.26E-01$	0.9781
16	545	$6.2500E-02$	$3.23E-03$	1.9687	$1.14E-01$	0.9929

Table 7.85: h-convergence for $\sigma_D = 1$, $\gamma = \gamma_3$, $Q = 1$, $d = 1$, $r = 1$, using $k = 1.0$. Errors e_{comb}

M	ndof	h	e_{comb}	EOC	residual
1	2	$1.0000E+00$	$2.57E+00$	-	4.9728e-16
2	3	$5.0000E-01$	$5.85E-01$	2.1385	9.5769e-16
4	5	$2.5000E-01$	$1.68E-01$	1.7991	3.6288e-15
8	9	$1.2500E-01$	$4.12E-02$	2.0261	7.2168e-16
16	17	$6.2500E-02$	$1.04E-02$	1.9906	5.8081e-15
32	33	$3.1250E-02$	$2.60E-03$	1.9963	1.2240e-14
64	65	$1.5625E-02$	$6.51E-04$	1.9973	2.3331e-14
128	129	$7.8125E-03$	$1.63E-04$	1.9986	7.9554e-14
256	257	$3.9062E-03$	$4.08E-05$	1.9993	5.3762e-11
512	513	$1.9531E-03$	$1.10E-05$	1.8895	9.7464e-08
1024	1025	$9.7656E-04$	$3.48E-06$	1.6600	2.6528e-07
2048	2049	$4.8828E-04$	$1.36E-06$	1.3583	7.9741e-07
4096	4097	$2.4414E-04$	$7.52E-07$	0.8539	1.5434e-06
8192	8193	$1.2207E-04$	$4.78E-07$	0.6514	3.0920e-06

Table 7.86: h-convergence for $\sigma_D = 1$, $\gamma = \gamma_3$, $Q = 1$, $d = 1$, $r = 1$, using $k = 1.0$. Errors e_0 and e_1

M	ndof	h	e_0	EOC	e_1	EOC
1	2	$1.0000E+00$	$5.87E-01$	-	$1.99E+00$	-
2	3	$5.0000E-01$	$1.61E-01$	1.8690	$8.48E-01$	1.2287
4	5	$2.5000E-01$	$5.71E-02$	1.4932	$4.44E-01$	0.9343
8	9	$1.2500E-01$	$1.32E-02$	2.1110	$2.24E-01$	0.9842
16	17	$6.2500E-02$	$3.35E-03$	1.9789	$1.12E-01$	0.9962
32	33	$3.1250E-02$	$8.43E-04$	1.9905	$5.62E-02$	0.9991
64	65	$1.5625E-02$	$2.12E-04$	1.9922	$2.81E-02$	0.9998
128	129	$7.8125E-03$	$5.31E-05$	1.9959	$1.41E-02$	0.9999
256	257	$3.9062E-03$	$1.33E-05$	1.9980	$7.03E-03$	1.0000
512	513	$1.9531E-03$	$4.14E-06$	1.6849	$3.52E-03$	1.0000
1024	1025	$9.7656E-04$	$1.77E-06$	1.2286	$1.76E-03$	1.0000
2048	2049	$4.8828E-04$	$9.29E-07$	0.9263	$8.79E-04$	1.0000
4096	4097	$2.4414E-04$	$6.44E-07$	0.5283	$4.40E-04$	0.9999
8192	8193	$1.2207E-04$	$4.52E-07$	0.5124	$2.20E-04$	0.9994

Table 7.87: h-convergence for $\sigma_D = 1$, $\gamma = \gamma_3$, $Q = 1$, $d = 3$, $r = 1$, using $k = 1.0$. Errors e_{comb}

M	ndof	h	e_{comb}	EOC	residual
1	8	$1.7321E+00$	$3.54E+00$	-	5.7468e-15
2	27	$8.6603E-01$	$1.55E+00$	1.1946	2.3364e-16
4	125	$4.3301E-01$	$5.51E-01$	1.4886	2.5884e-16
8	729	$2.1651E-01$	$1.55E-01$	1.8286	4.6729e-16
16	4913	$1.0825E-01$	$4.05E-02$	1.9370	5.9415e-16
32	35937	$5.4127E-02$	$1.03E-02$	1.9796	8.5064e-16

Table 7.88: h-convergence for $\sigma_D = 1$, $\gamma = \gamma_3$, $Q = 1$, $d = 3$, $r = 1$, using $k = 1.0$. Errors e_0 and e_1

M	ndof	h	e_0	EOC	e_1	EOC
1	8	$1.7321E+00$	$3.65E-01$	-	$1.83E+00$	-
2	27	$8.6603E-01$	$1.59E-01$	1.1999	$1.60E+00$	0.1940
4	125	$4.3301E-01$	$1.01E-01$	0.6554	$1.04E+00$	0.6236
8	729	$2.1651E-01$	$3.25E-02$	1.6331	$5.67E-01$	0.8762
16	4913	$1.0825E-01$	$8.84E-03$	1.8772	$2.93E-01$	0.9533
32	35937	$5.4127E-02$	$2.27E-03$	1.9607	$1.48E-01$	0.9849

Table 7.89: h-convergence for $\sigma_D = 1$, $\gamma = \gamma_3$, $Q = 3$, $d = 1$, $r = 2$, using $k = 1.0$. Errors e_{comb}

M	ndof	h	e_{comb}	EOC	residual
1	3	$1.0000E+00$	$3.67E-01$	-	6.6613e-16
2	5	$5.0000E-01$	$1.58E-01$	1.2169	1.4390e-15
4	9	$2.5000E-01$	$4.22E-02$	1.9040	1.0208e-14
8	17	$1.2500E-01$	$1.02E-02$	2.0511	2.1488e-14
16	33	$6.2500E-02$	$1.54E-03$	2.7299	1.3609e-13
32	65	$3.1250E-02$	$2.20E-04$	2.8062	6.0838e-13
64	129	$1.5625E-02$	$1.92E-05$	3.5167	1.8526e-12
128	257	$7.8125E-03$	$1.72E-06$	3.4758	7.6150e-12
256	513	$3.9062E-03$	$2.00E-07$	3.1066	1.6222e-10
512	1025	$1.9531E-03$	$1.65E-08$	3.6027	1.9903e-09
1024	2049	$9.7656E-04$	$2.05E-09$	3.0090	4.3915e-10
2048	4097	$4.8828E-04$	$4.06E-10$	2.3348	3.0360e-09
4096	8193	$2.4414E-04$	$2.42E-09$	-2.5772	8.3669e-09

Table 7.90: h-convergence for $\sigma_D = 1$, $\gamma = \gamma_3$, $Q = 3$, $d = 1$, $r = 2$, using $k = 1.0$. Errors e_0 and e_1

M	ndof	h	e_0	EOC	e_1	EOC
1	3	$1.0000E+00$	$9.55E-02$	-	$2.72E-01$	-
2	5	$5.0000E-01$	$5.83E-02$	0.7110	$1.99E-01$	0.4471
4	9	$2.5000E-01$	$2.93E-02$	0.9913	$5.15E-02$	1.9521
8	17	$1.2500E-01$	$8.57E-03$	1.7745	$1.29E-02$	1.9987
16	33	$6.2500E-02$	$1.33E-03$	2.6847	$3.23E-03$	1.9978
32	65	$3.1250E-02$	$1.94E-04$	2.7791	$8.07E-04$	1.9994
64	129	$1.5625E-02$	$1.60E-05$	3.5998	$2.02E-04$	1.9998
128	257	$7.8125E-03$	$1.33E-06$	3.5912	$5.05E-05$	2.0000
256	513	$3.9062E-03$	$1.51E-07$	3.1398	$1.26E-05$	2.0000
512	1025	$1.9531E-03$	$1.03E-08$	3.8704	$3.15E-06$	2.0000
1024	2049	$9.7656E-04$	$1.28E-09$	3.0143	$7.88E-07$	2.0000
2048	4097	$4.8828E-04$	$3.09E-10$	2.0445	$1.97E-07$	2.0000
4096	8193	$2.4414E-04$	$2.41E-09$	-2.9607	$4.94E-08$	1.9971

Table 7.91: h-convergence for $\sigma_D = 1$, $\gamma = \gamma_3$, $Q = 3$, $d = 1$, $r = 3$, using $k = 1.0$. Errors e_{comb}

M	ndof	h	e_{comb}	EOC	residual
1	4	$1.0000E+00$	$3.81E-01$	-	2.1065e-15
2	7	$5.0000E-01$	$3.96E-02$	3.2681	4.2798e-15
4	13	$2.5000E-01$	$1.64E-02$	1.2749	3.7812e-14
8	25	$1.2500E-01$	$1.63E-03$	3.3281	8.6197e-14
16	49	$6.2500E-02$	$2.01E-04$	3.0219	4.2647e-13
32	97	$3.1250E-02$	$2.30E-05$	3.1244	1.5107e-12
64	193	$1.5625E-02$	$1.45E-06$	3.9855	6.0228e-12
128	385	$7.8125E-03$	$2.39E-07$	2.6037	1.7047e-11
256	769	$3.9062E-03$	$1.30E-08$	4.1961	1.0353e-10
512	1537	$1.9531E-03$	$2.05E-09$	2.6704	1.0748e-09
1024	3073	$9.7656E-04$	$2.00E-09$	0.0316	3.8353e-09
2048	6145	$4.8828E-04$	$7.18E-09$	-1.8412	1.5013e-08

Table 7.92: h-convergence for $\sigma_D = 1$, $\gamma = \gamma_3$, $Q = 3$, $d = 1$, $r = 3$, using $k = 1.0$. Errors e_0 and e_1

M	ndof	h	e_0	EOC	e_1	EOC
1	4	$1.0000E+00$	$1.10E-01$	-	$2.71E-01$	-
2	7	$5.0000E-01$	$2.61E-02$	2.0727	$2.69E-02$	3.3357
4	13	$2.5000E-01$	$1.55E-02$	0.7567	$3.54E-03$	2.9250
8	25	$1.2500E-01$	$1.57E-03$	3.2984	$4.48E-04$	2.9820
16	49	$6.2500E-02$	$1.97E-04$	2.9962	$5.46E-05$	3.0352
32	97	$3.1250E-02$	$2.28E-05$	3.1128	$6.71E-06$	3.0254
64	193	$1.5625E-02$	$1.44E-06$	3.9854	$8.37E-07$	3.0036
128	385	$7.8125E-03$	$2.38E-07$	2.5956	$1.05E-07$	2.9999
256	769	$3.9062E-03$	$1.30E-08$	4.1968	$1.31E-08$	3.0002
512	1537	$1.9531E-03$	$2.04E-09$	2.6671	$1.70E-09$	2.9427
1024	3073	$9.7656E-04$	$2.00E-09$	0.0307	$2.09E-09$	-0.3007
2048	6145	$4.8828E-04$	$7.17E-09$	-1.8419	$8.05E-09$	-1.9414

Table 7.93: h-convergence for $\sigma_D = 1$, $\gamma = \gamma_3$, $Q = 3$, $d = 1$, $r = 4$, using $k = 1.0$. Errors e_{comb}

M	ndof	h	e_{comb}	EOC	residual
1	5	$1.0000E+00$	$1.40E-01$	-	2.9458e-15
2	9	$5.0000E-01$	$4.53E-02$	1.6235	1.4220e-14
4	17	$2.5000E-01$	$9.74E-03$	2.2196	1.0249e-13
8	33	$1.2500E-01$	$4.21E-04$	4.5304	3.2550e-13
16	65	$6.2500E-02$	$5.50E-05$	2.9379	1.2463e-12
32	129	$3.1250E-02$	$8.20E-06$	2.7445	6.1113e-12
64	257	$1.5625E-02$	$5.47E-07$	3.9075	1.9428e-11
128	513	$7.8125E-03$	$4.55E-08$	3.5878	1.4315e-10
256	1025	$3.9062E-03$	$6.27E-09$	2.8573	1.9188e-09
512	2049	$1.9531E-03$	$3.19E-09$	0.9763	3.6601e-09
1024	4097	$9.7656E-04$	$1.16E-08$	-1.8570	2.6101e-08
2048	8193	$4.8828E-04$	$3.77E-08$	-1.7081	1.5876e-07

Table 7.94: h-convergence for $\sigma_D = 1$, $\gamma = \gamma_3$, $Q = 3$, $d = 1$, $r = 4$, using $k = 1.0$. Errors e_0 and e_1

M	ndof	h	e_0	EOC	e_1	EOC
1	5	$1.0000E+00$	$1.27E-01$	-	$1.28E-02$	-
2	9	$5.0000E-01$	$4.24E-02$	1.5812	$5.83E-03$	1.1300
4	17	$2.5000E-01$	$9.58E-03$	2.1469	$6.19E-04$	3.2333
8	33	$1.2500E-01$	$4.20E-04$	4.5123	$1.17E-05$	5.7301
16	65	$6.2500E-02$	$5.49E-05$	2.9343	$8.39E-07$	3.7988
32	129	$3.1250E-02$	$8.20E-06$	2.7440	$1.62E-07$	2.3743
64	257	$1.5625E-02$	$5.47E-07$	3.9070	$9.59E-09$	4.0767
128	513	$7.8125E-03$	$4.55E-08$	3.5874	$3.17E-10$	4.9168
256	1025	$3.9062E-03$	$6.27E-09$	2.8578	$6.55E-10$	-1.0450
512	2049	$1.9531E-03$	$3.18E-09$	0.9785	$3.10E-09$	-2.2453
1024	4097	$9.7656E-04$	$1.15E-08$	-1.8583	$1.17E-08$	-1.9173
2048	8193	$4.8828E-04$	$3.77E-08$	-1.7087	$4.28E-08$	-1.8682

Table 7.95: h-convergence for $\sigma_D = 1$, $\gamma = \gamma_3$, $Q = 4$, $d = 1$, $r = 1$, using $k = 0.0001$. Errors e_{comb}

M	ndof	h	e_{comb}	EOC	residual
1	2	$1.0000E+00$	$2.77E+00$	-	3.6663e-12
2	3	$5.0000E-01$	$5.83E-01$	2.2476	4.6431e-12
4	5	$2.5000E-01$	$1.88E-01$	1.6323	2.8125e-12
8	9	$1.2500E-01$	$1.17E-01$	0.6808	1.2962e-12
16	17	$6.2500E-02$	$6.52E-02$	0.8477	8.3215e-13
32	33	$3.1250E-02$	$2.75E-02$	1.2449	4.8385e-13
64	65	$1.5625E-02$	$7.34E-03$	1.9054	6.1056e-13
128	129	$7.8125E-03$	$1.83E-03$	2.0009	6.0372e-13

Table 7.96: h-convergence for $\sigma_D = 1$, $\gamma = \gamma_3$, $Q = 4$, $d = 1$, $r = 1$, using $k = 0.0001$. Errors e_0 and e_1

M	ndof	h	e_0	EOC	e_1	EOC
1	2	$1.0000E+00$	$7.60E-01$	-	$2.01E+00$	-
2	3	$5.0000E-01$	$1.57E-01$	2.2780	$8.53E-01$	1.2363
4	5	$2.5000E-01$	$7.66E-02$	1.0321	$4.46E-01$	0.9357
8	9	$1.2500E-01$	$8.91E-02$	-0.2178	$2.26E-01$	0.9826
16	17	$6.2500E-02$	$5.81E-02$	0.6168	$1.13E-01$	0.9957
32	33	$3.1250E-02$	$2.57E-02$	1.1751	$5.65E-02$	1.0007
64	65	$1.5625E-02$	$6.90E-03$	1.8991	$2.82E-02$	1.0015
128	129	$7.8125E-03$	$1.72E-03$	2.0009	$1.41E-02$	1.0005

Table 7.97: h-convergence for $\sigma_D = 1$, $\gamma = \gamma_3$, $Q = 4$, $d = 1$, $r = 1$, using $k = 0.25$. Errors e_{comb}

M	ndof	h	e_{comb}	EOC	residual
1	2	$1.0000E+00$	$2.75E+00$	-	1.1102e-15
2	3	$5.0000E-01$	$5.82E-01$	2.2418	5.1692e-14
4	5	$2.5000E-01$	$1.74E-01$	1.7417	7.2542e-14
8	9	$1.2500E-01$	$1.17E-01$	0.5686	6.9873e-14
16	17	$6.2500E-02$	$1.10E-01$	0.0927	9.0977e-14
32	33	$3.1250E-02$	$1.18E-01$	-0.1055	9.8859e-14
64	65	$1.5625E-02$	$1.16E-01$	0.0254	1.4237e-13
128	129	$7.8125E-03$	$1.09E-01$	0.0884	3.5941e-13

Table 7.98: h-convergence for $\sigma_D = 1$, $\gamma = \gamma_3$, $Q = 4$, $d = 1$, $r = 1$, using $k = 0.25$. Errors e_0 and e_1

M	ndof	h	e_0	EOC	e_1	EOC
1	2	$1.0000E+00$	$7.50E-01$	-	$2.00E+00$	-
2	3	$5.0000E-01$	$1.56E-01$	2.2623	$8.52E-01$	1.2342
4	5	$2.5000E-01$	$6.27E-02$	1.3186	$4.46E-01$	0.9346
8	9	$1.2500E-01$	$8.92E-02$	-0.5085	$2.25E-01$	0.9825
16	17	$6.2500E-02$	$1.03E-01$	-0.2079	$1.13E-01$	0.9987
32	33	$3.1250E-02$	$1.17E-01$	-0.1793	$5.66E-02$	0.9956
64	65	$1.5625E-02$	$1.16E-01$	0.0092	$2.84E-02$	0.9952
128	129	$7.8125E-03$	$1.09E-01$	0.0843	$1.43E-02$	0.9849

Table 7.99: h-convergence for $\sigma_D = 1$, $\gamma = \gamma_3$, $Q = 5$, $d = 1$, $r = 1$, using $k = 0.0001$. Errors e_{comb}

M	ndof	h	e_{comb}	EOC	residual
1	2	$1.0000E+00$	$2.77E+00$	-	3.6663e-12
2	3	$5.0000E-01$	$5.83E-01$	2.2476	5.5556e-12
4	5	$2.5000E-01$	$1.88E-01$	1.6323	2.4699e-12
8	9	$1.2500E-01$	$1.17E-01$	0.6808	1.2297e-12
16	17	$6.2500E-02$	$6.52E-02$	0.8477	8.0200e-13
32	33	$3.1250E-02$	$2.75E-02$	1.2449	4.8719e-13
64	65	$1.5625E-02$	$7.34E-03$	1.9054	6.5180e-13
128	129	$7.8125E-03$	$1.83E-03$	2.0009	6.2732e-13

Table 7.100: h-convergence for $\sigma_D = 1$, $\gamma = \gamma_3$, $Q = 5$, $d = 1$, $r = 1$, using $k = 0.0001$. Errors e_0 and e_1

M	ndof	h	e_0	EOC	e_1	EOC
1	2	$1.0000E+00$	$7.60E-01$	-	$2.01E+00$	-
2	3	$5.0000E-01$	$1.57E-01$	2.2780	$8.53E-01$	1.2363
4	5	$2.5000E-01$	$7.66E-02$	1.0321	$4.46E-01$	0.9357
8	9	$1.2500E-01$	$8.91E-02$	-0.2178	$2.26E-01$	0.9826
16	17	$6.2500E-02$	$5.81E-02$	0.6168	$1.13E-01$	0.9957
32	33	$3.1250E-02$	$2.57E-02$	1.1751	$5.65E-02$	1.0007
64	65	$1.5625E-02$	$6.90E-03$	1.8991	$2.82E-02$	1.0015
128	129	$7.8125E-03$	$1.72E-03$	2.0009	$1.41E-02$	1.0005

Table 7.101: h-convergence for $\sigma_D = 1$, $\gamma = \gamma_3$, $Q = 5$, $d = 1$, $r = 1$, using $k = 0.25$. Errors e_{comb}

M	ndof	h	e_{comb}	EOC	residual
1	2	$1.0000E+00$	$2.75E+00$	-	1.1102e-15
2	3	$5.0000E-01$	$5.82E-01$	2.2432	9.1648e-14
4	5	$2.5000E-01$	$1.99E-01$	1.5442	9.8473e-14
8	9	$1.2500E-01$	$1.43E-01$	0.4845	9.7266e-14
16	17	$6.2500E-02$	$1.60E-01$	-0.1678	8.7611e-14
32	33	$3.1250E-02$	$1.78E-01$	-0.1541	9.1487e-14
64	65	$1.5625E-02$	$1.77E-01$	0.0083	1.4025e-13
128	129	$7.8125E-03$	$1.73E-01$	0.0350	3.8546e-13

Table 7.102: h-convergence for $\sigma_D = 1$, $\gamma = \gamma_3$, $Q = 5$, $d = 1$, $r = 1$, using $k = 0.25$. Errors e_0 and e_1

M	ndof	h	e_0	EOC	e_1	EOC
1	2	$1.0000E+00$	$7.50E-01$	-	$2.00E+00$	-
2	3	$5.0000E-01$	$1.56E-01$	2.2678	$8.52E-01$	1.2341
4	5	$2.5000E-01$	$8.80E-02$	0.8244	$4.46E-01$	0.9340
8	9	$1.2500E-01$	$1.14E-01$	-0.3775	$2.26E-01$	0.9806
16	17	$6.2500E-02$	$1.53E-01$	-0.4212	$1.13E-01$	0.9993
32	33	$3.1250E-02$	$1.76E-01$	-0.2046	$5.71E-02$	0.9853
64	65	$1.5625E-02$	$1.77E-01$	-0.0025	$2.91E-02$	0.9714
128	129	$7.8125E-03$	$1.73E-01$	0.0323	$1.55E-02$	0.9088

Table 7.103: k-convergence for $\sigma_D = 1$, $\gamma = \gamma_3$, $Q = 1$, $d = 2$, $r = 1$. Errors e_{comb}. Here $\hat{\gamma} = 10$.

N	k	e_{comb}	EOC	residual
1	$1.0000E+00$	$4.78E-01$	-	3.0987e-15
2	$5.0000E-01$	$3.79E-01$	0.3330	1.8116e-15
4	$2.5000E-01$	$1.39E-01$	1.4531	1.9984e-15
8	$1.2500E-01$	$3.53E-02$	1.9717	1.5223e-15
16	$6.2500E-02$	$8.13E-03$	2.1192	2.8219e-15
32	$3.1250E-02$	$1.89E-03$	2.1045	4.1928e-15
64	$1.5625E-02$	$4.46E-04$	2.0829	9.2303e-15
128	$7.8125E-03$	$1.07E-04$	2.0545	1.8115e-14
256	$3.9062E-03$	$2.63E-05$	2.0321	4.7266e-14
512	$1.9531E-03$	$6.49E-06$	2.0174	8.6464e-14
1024	$9.7656E-04$	$1.61E-06$	2.0091	1.8906e-13
2048	$4.8828E-04$	$4.02E-07$	2.0046	4.2916e-13

Table 7.104: k-convergence for $\sigma_D = 1$, $\gamma = \gamma_3$, $Q = 1$, $d = 2$, $r = 1$. Errors e_0 and e_1. Here $\hat{\gamma} = 10$.

N	k	e_0	EOC	e_1	EOC
1	$1.0000E+00$	$4.32E-01$	-	$4.54E-02$	-
2	$5.0000E-01$	$3.26E-01$	0.4074	$5.33E-02$	-0.2314
4	$2.5000E-01$	$1.17E-01$	1.4835	$2.20E-02$	1.2797
8	$1.2500E-01$	$2.96E-02$	1.9770	$5.71E-03$	1.9438
16	$6.2500E-02$	$6.82E-03$	2.1190	$1.31E-03$	2.1204
32	$3.1250E-02$	$1.59E-03$	2.1044	$3.05E-04$	2.1052
64	$1.5625E-02$	$3.74E-04$	2.0831	$7.21E-05$	2.0816
128	$7.8125E-03$	$9.01E-05$	2.0548	$1.74E-05$	2.0533
256	$3.9062E-03$	$2.20E-05$	2.0322	$4.25E-06$	2.0313
512	$1.9531E-03$	$5.44E-06$	2.0175	$1.05E-06$	2.0170
1024	$9.7656E-04$	$1.35E-06$	2.0091	$2.61E-07$	2.0089
2048	$4.8828E-04$	$3.37E-07$	2.0047	$6.50E-08$	2.0045

7.5 Behaviour of the solution

In this section we present some numerical examples to show what type of behaviour the solution of the Cohen model can have.

In many cases the solution for u develops a *sharp front*, a big change in value, almost a discontinuity. To describe this behaviour we define $x_m(t)$ as the position of the largest gradient in one-dimensional domains, i.e. the value $x_m(t) \in \Omega$ such that $|\nabla u(t, x_m(t))|$ is maximised. To get a similar description for 2D and 3D domains, it seems natural to locate the largest gradient along a line that crosses the domain.

A question that has been raised, see for example [Edw95], is how the position of the sharp front changes with time. One possibility is that it settles down at a fixed position. We will see some examples of this behaviour in this section, see Subsections 7.5.1, 7.5.4, 7.5.5 and 7.5.6.

If the domain is $\Omega = (0, \infty)$ then different behaviour is possible, see [NLM$^+$] and Subsection 7.5.2: The sharp front can move such that $x_m(t) = Ct^{1/2}$.

Other results show that the wave front can also travel at constant speed, see for example [SE04].

On bounded domains, however, it appears as if the sharp front will always become stationary. Although it is possible to obtain solutions where a wave front is moving at constant speed, this wave front is actually not as sharp as in the stationary case. Moreover it seems as if this wave front is not due to the nonlinearity but to a similarity of the Cohen model to the wave equation, see Subsection 7.5.3.

Another aspect of the behaviour of the solution is the so-called *mass uptake* $M_u(t)$, which is defined by

$$M_u(t) := \int_\Omega u(t, \boldsymbol{x})\, d\boldsymbol{x}. \tag{7.7}$$

Unlike the heat equation, the Cohen model allows for solutions where the mass uptake can first increase and then decrease again. This behaviour is called *sorption overshoot*. The wave-like solution of Subsection 7.5.3 is an extreme case of this behaviour.

Unless specified otherwise, the following values were chosen for the parameters: $D = K = \mu = 1$ (and $\nu = 0$). Also, in all the examples of this section, no Dirichlet boundary conditions are prescribed for σ, i.e. $\sigma_D = 0$. The main reason for this is that in these examples the Dirichlet conditions for u are inhomogeneous, therefore (in order to satisfy (3.1b) on Γ_D) the Dirichlet conditions for σ would have to be calculated.

7.5.1 Sharp front in 1D

We now bring a one-dimensional example where the solution develops a sharp front. The domain is $\Omega = (0, 1)$ with $\Gamma_D = \partial\Omega = \{0, 1\}$. The Dirichlet boundary conditions are $u_D = 1 - \exp(-t/2)$ for $x = 0$ and $u_D = 0$ for $x = 1$. The other parameters of the problem are $\gamma = \gamma_3$, $\check{\gamma} = 0.5$, $\hat{\gamma} = 1.0$, $u_a = 0.5$, $\Delta_a = 0.05$ and $T = 30$. The discretisation parameters are $Q = 3$, $k = 0.015$, $h = 1/256$ and $r = 3$. See Figure 7.2 for plots of the numerical solution. It can be seen that u develops a sharp front, which remains stationary as can be seen from the plot of $x_m(t)$. Furthermore σ has a sharp front in the opposite direction so that the sum $u + \sigma$ has no sharp front at all. The mass uptake does not show sorption overshoot behaviour.

7.5.2 Infinite domain in 1D: $\Omega = (0, \infty)$

In this subsection we present some numerical experiments that were made to support theoretical results by Needham et al in [NLM$^+$]. In order to see that the numerical experiments support the theory, we first state the theoretical results.

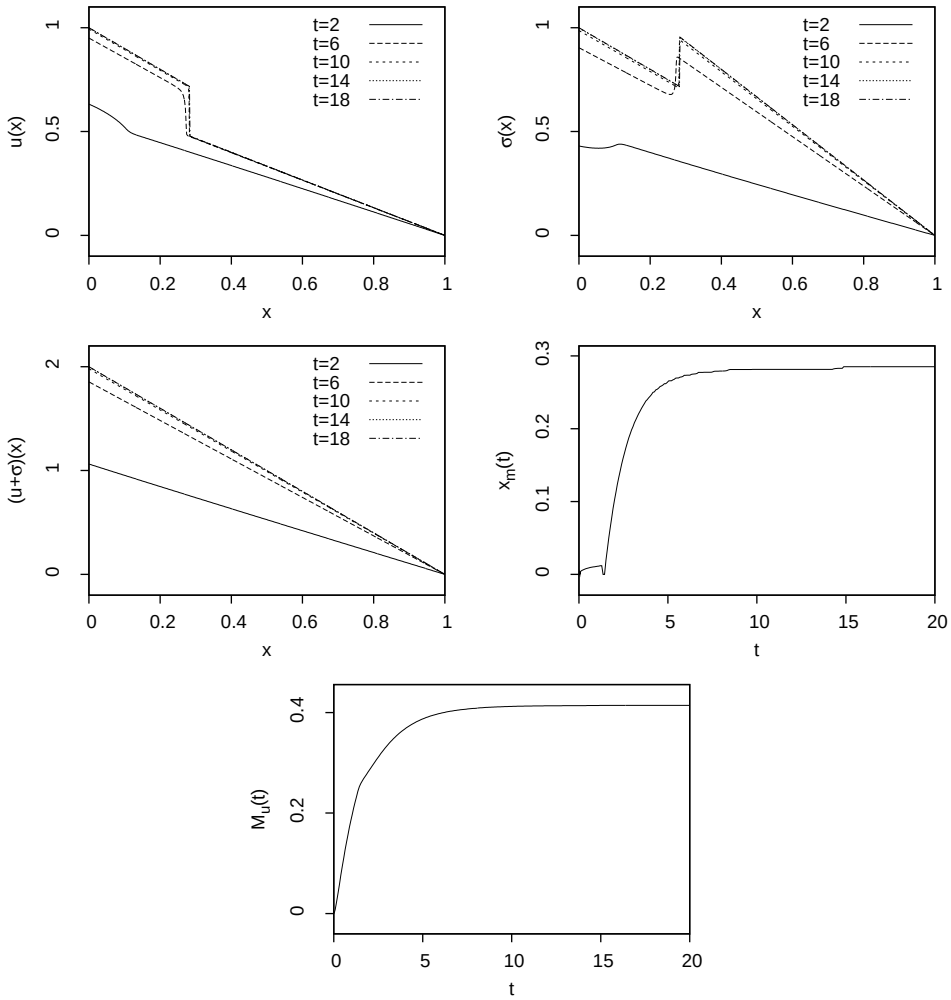

Figure 7.2: Numerical solution on $\Omega = (0, 1)$.

Equations

The equations considered in [NLM$^+$] are based on Problem 3.1 but use a different scaling: instead of $\mu = 1$, this scaling gives $\check{\gamma} = 1$. The domain here is unbounded and one-dimensional:

$$\dot{u} = \Delta(u + \sigma) \quad \text{for } (x,t) \in (0,\infty) \times [0,T], \tag{7.8a}$$

$$\dot{\sigma} + \gamma(u)\sigma = \mu u \quad \text{for } (x,t) \in (0,\infty) \times [0,T], \tag{7.8b}$$

with $u(t,0) = u_s$ and a uniform decay condition as $x \to \infty$.

Defining functions G, D

Omitting the $\dot{\sigma}$-term from (7.8b) would give

$$\sigma = \frac{\mu u}{\gamma(u)},$$

and using this in (7.8a), we obtain

$$\dot{u} = \Delta G(u) = \nabla^T D(u) \nabla u,$$

where

$$\begin{aligned} G(u) &:= u + \frac{\mu u}{\gamma(u)}, \\ D(u) &:= G'(u) = 1 + \frac{\mu(\gamma(u) - u\gamma'(u))}{\gamma(u)^2}. \end{aligned} \tag{7.9}$$

The three types and their subtypes

Since $D(u)$ can become negative for certain values of the parameters, there are three types to consider:

- Type I: $\min_{u>0} D(u) > 0$,
- Type II: $\min_{u>0} D(u) = 0$,
- Type III: $\min_{u>0} D(u) < 0$.

For fixed values of μ and $\hat{\gamma}$, there seems to be a unique critical value of Δ, which we will denote by Δ^*, such that

- $\Delta > \Delta^* \quad \Rightarrow \quad$ Type I,
- $\Delta = \Delta^* \quad \Rightarrow \quad$ Type II,
- $\Delta < \Delta^* \quad \Rightarrow \quad$ Type III.

Type I. See Figure 7.3 for typical plots of functions D and G. This type is like classical diffusion, no sharp front is developing.

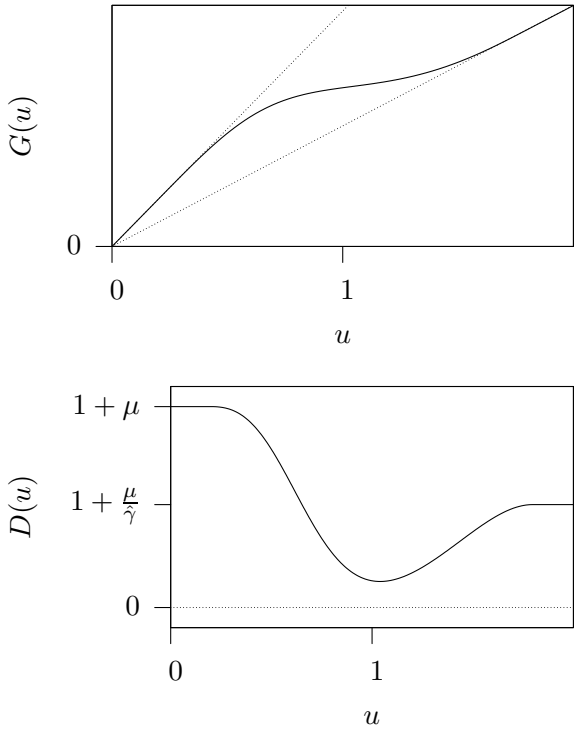

Figure 7.3: Typical shape of functions G and D for Type I.

Type II. See Figure 7.4 for typical plots of functions D and G for Type II. There seems to be a unique value of $u > 0$, denoted by u^*, for which $D(u)$ achieves its minimum, i.e. such that $D(u^*) = 0$. There are two subtypes to consider, depending on u_s, the value of u at $x = 0$:

- Type IIa: $u_s < u^*$.
  Like classical diffusion, no sharp front is developing.
- Type IId: $u_s > u^*$.
  Moving sharp front develops at $x_m(t) = Ct^{1/2}$.

Note that we call the second subtype IId here (and not IIb) to emphasise the similarity to subtype IIId, see the next paragraph.

Type III. For this type, the equation $D(u) = 0$ has two positive solutions, compare Figure 7.5. Define u_- and u_+ via

$$D(u_-) = D(u_+) = 0 \text{ and } 0 < u_- < u_+.$$

Define also u_c via

$$G(u_c) = G(u_-) \text{ and } u_c > u_+.$$

Thus

$$0 < u_- < u_+ < u_c.$$

There are four subtypes to consider, again depending on u_s:

- Type IIIa: $0 < u_s < u_-$.
  Like classical diffusion, no sharp front is developing.

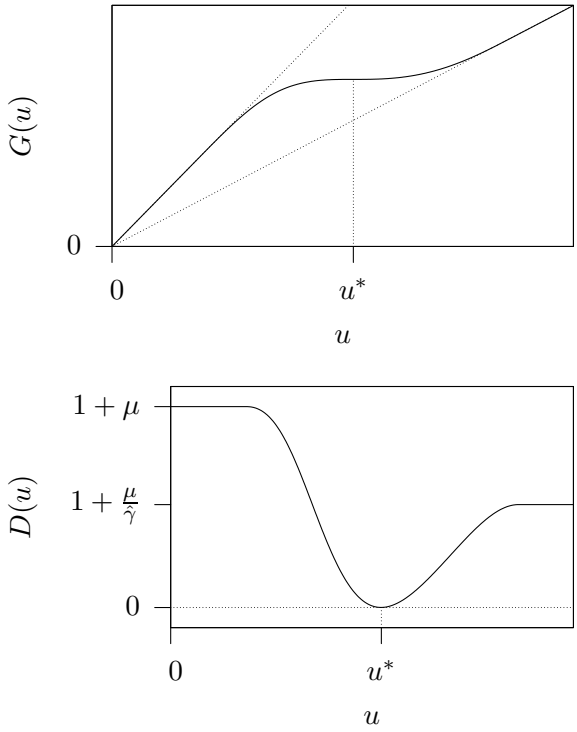

Figure 7.4: Typical shape of functions G and D for Type II.

- Type IIIb: $u_- < u_s < u_+$.
  Stationary sharp front develops at $x_m(t) = 0$.
- Type IIIc: $u_+ < u_s < u_c$.
  Stationary sharp front develops at $x_m(t) = C > 0$.
- Type IIId: $u_c < u_s < \infty$.
  Moving sharp front develops at $x_m(t) = Ct^{1/2}$.

For convenience we summarise the results predicted by theory in Table 7.105.

Table 7.105: Infinite domain in 1D: Summary of types

Type	**Behaviour of u from theory**	**Numerical example**
I	like classical diffusion	Figure 7.7
IIa	like classical diffusion	Figure 7.10
IId	moving sharp front at $x_m(t) = Ct^{1/2}$	Figure 7.11
IIIa	like classical diffusion	Figure 7.14
IIIb	stationary sharp front at $x_m(t) = 0$	Figure 7.15
IIIc	stationary sharp front at $x_m(t) = C > 0$	Figure 7.16
IIId	moving sharp front at $x_m(t) = Ct^{1/2}$	Figure 7.17

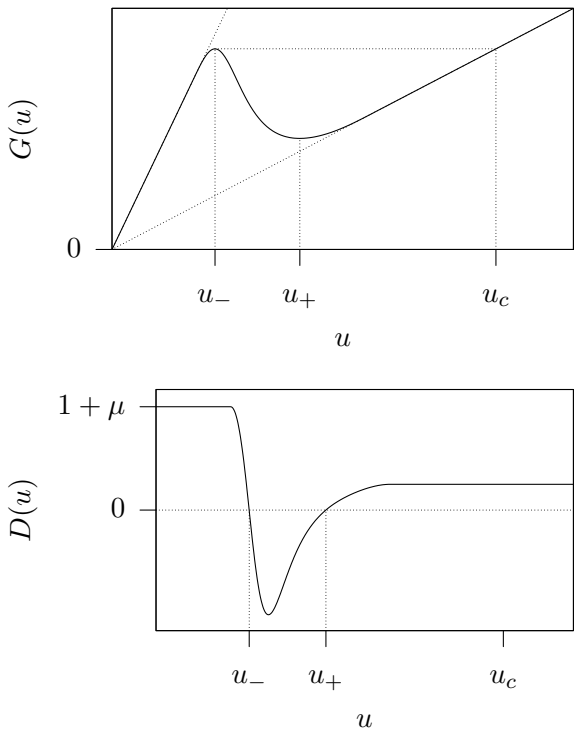

Figure 7.5: Typical shape of functions G and D for Type III.

About the plots

As an example, we present the results for the values $\gamma = \gamma_3$, $\hat{\gamma} = 2$ and $\mu = 1$. For these values we determine the value of Δ^* for Type II and the corresponding value u^*. Using these, the values of Δ and u_s are chosen as follows:

- Type I: $\Delta = 2\Delta^*$ and $u_s = 10$.
- Type II: $\Delta = \Delta^*$.
  - Type IIa: $u_s = \frac{1}{2}u^*$.
  - Type IId: $u_s = 2u^*$.
- Type III: $\Delta = \frac{1}{2}\Delta^*$. For this value of Δ determine u_-, u_+ and u_c.
  - Type IIIa: $u_s = \frac{1}{2}u_-$.
  - Type IIIb: $u_s = \frac{1}{2}(u_- + u_+)$.
  - Type IIIc: $u_s = \frac{1}{2}(u_+ + u_c)$.
  - Type IIId: $u_s = 2u_c$.

The next pages contain the following figures:

- Plots of $\gamma(u)$, $\gamma'(u)$, $\gamma''(u)$, $G(u)$ and $D(u)$ for each of the three types I, II and III, see Figures 7.6, 7.8 and 7.12. For types II and III an enlarged version of $G(u)$ and $D(u)$ is included, which shows the critical values of u, see Figures 7.9 and 7.13.

- Plots of the numerical solution $u(x)$, $\sigma(x)$, $(u+\sigma)(x)$ for each of the full types I, IIa, IId, IIIa, IIIb, IIIc, IIId. For types IId, IIIb, IIIc, IIId a plot of the position $x_m(t)$ of the steepest gradient of $u(x)$ is given as well. See Figure 7.7, Figures 7.10-7.11 and Figures 7.14-7.17.

For all of these simulations we take $T = 20$ as the final time and use 2000 time steps. To simulate $x \in (0, \infty)$ we take the spatial domain to be $\Omega = (0, 60)$. The uniform decay condition for u as $x \to \infty$ is simulated by enforcing a homogeneous Dirichlet boundary condition at the far end of the domain, i.e. at $x = 60$. The space mesh is refined several times in a right-hand neighbourhood of $x = 0$ (and near $x_m(t = 20)$ for Type IIIc).

By comparing the numerical results with the theoretical results in Table 7.105 it can be seen that both agree.

About finding the critical values

For fixed values of μ and $\hat{\gamma}$ the critical values Δ^* and u^* are found using the following method.

Define the function $M : \mathbb{R}_{>0} \to \mathbb{R}$ via

$$M(\Delta) = \min_{u>0} D(u; \Delta).$$

Calculating $M(\Delta)$ involved these steps:

- Divide the interval $[1 - 2\Delta, 1 + 2\Delta] \cap \mathbb{R}_{\geq 0}$ into 100 subintervals with endpoints u_i, $i = 0, \ldots, 100$. Locate a bracket $[u_{i-1}, u_{i+1}]$ containing the minimum by minimising $D(u_i)$.

- Find $u_{\min}$, the position of the minimum of D, by applying a root-finding method to
$$D'_f(u) = u\gamma(u)\gamma''(u) + 2\gamma(u)\gamma'(u) - 2u\gamma'(u)^2$$
in the bracket $[u_{i-1}, u_{i+1}]$ with tolerance 10^{-15}. Note that D'_f is a factor of D' such that $D'_f(u) = 0 \iff D'(u) = 0$.

- Set $M(\Delta) = D(u_{\min})$.

The critical value Δ^* can now be found by applying a root-finding method to function M with tolerance 10^{-12}. For u^*, we take the value of $u_{\min}$ corresponding to Δ^*.

The values of u_-, u_+ and u_c are also found using a root-finding method, since they are all just solutions to one-dimensional nonlinear equations.

The root-finding method used throughout was Brent's method, see [PTVF02, Section 9.3].

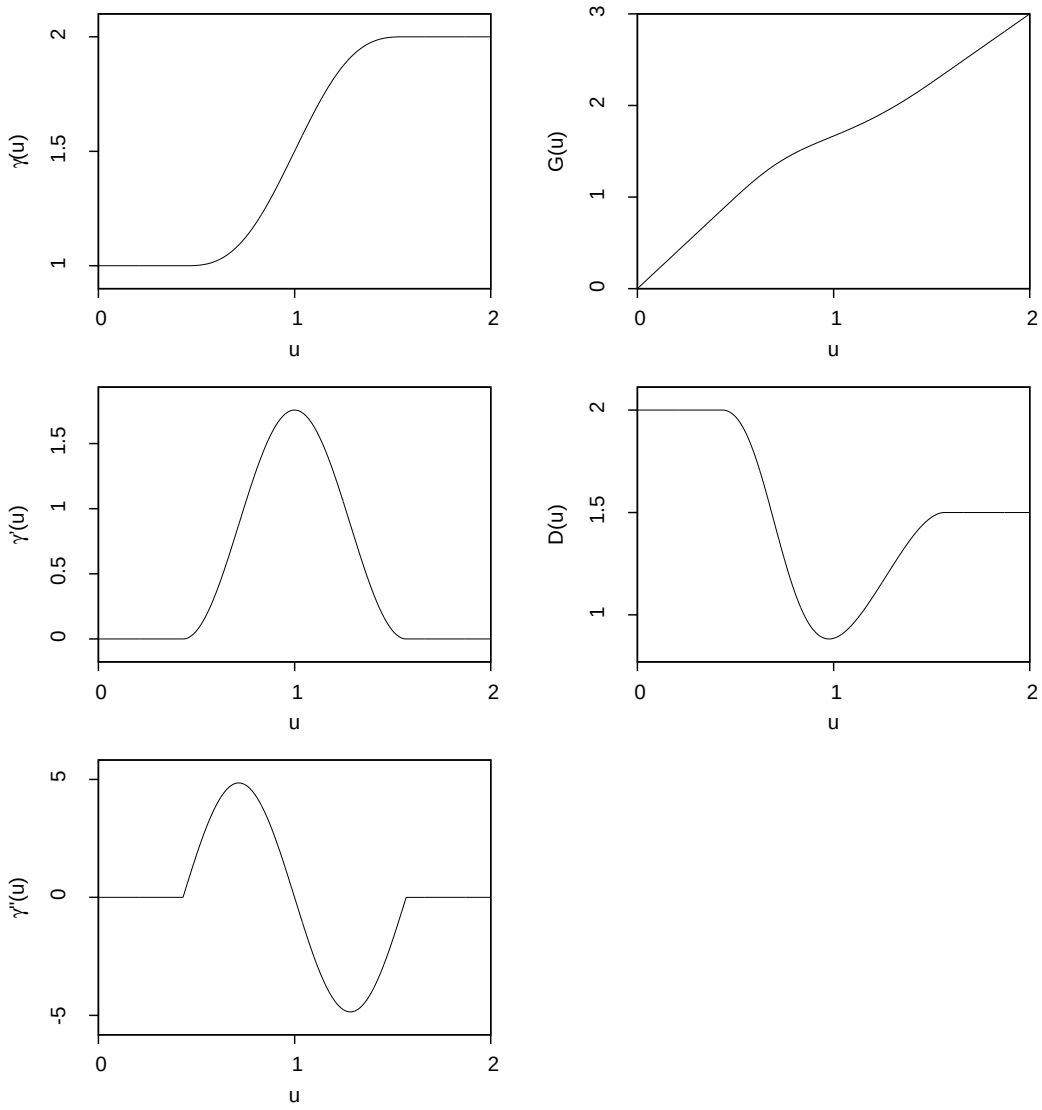

Figure 7.6: Plot of $\gamma(u)$, $\gamma'(u)$, $\gamma''(u)$, $G(u)$ and $D(u)$, Type I, $\Delta_a = 1.13786 > \Delta^*$.

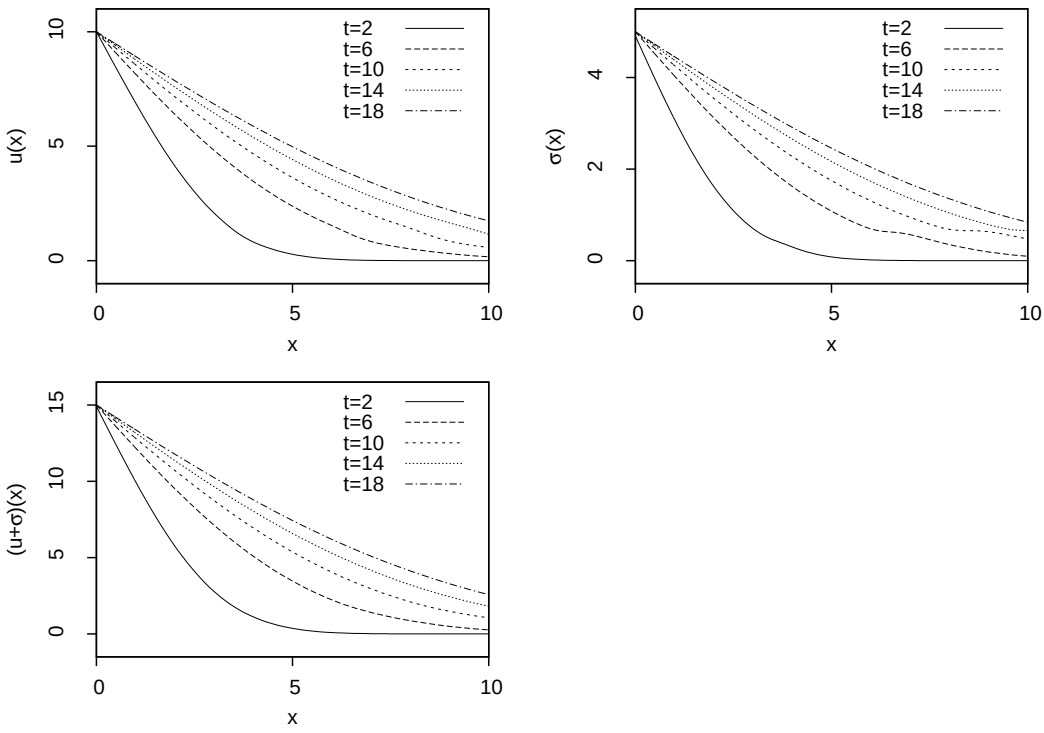

Figure 7.7: Numerical solution $u(x)$, $\sigma(x)$, $(u+\sigma)(x)$ for Type I, $\Delta_a = 1.13786 > \Delta^*$. Also $u_s = 10$.

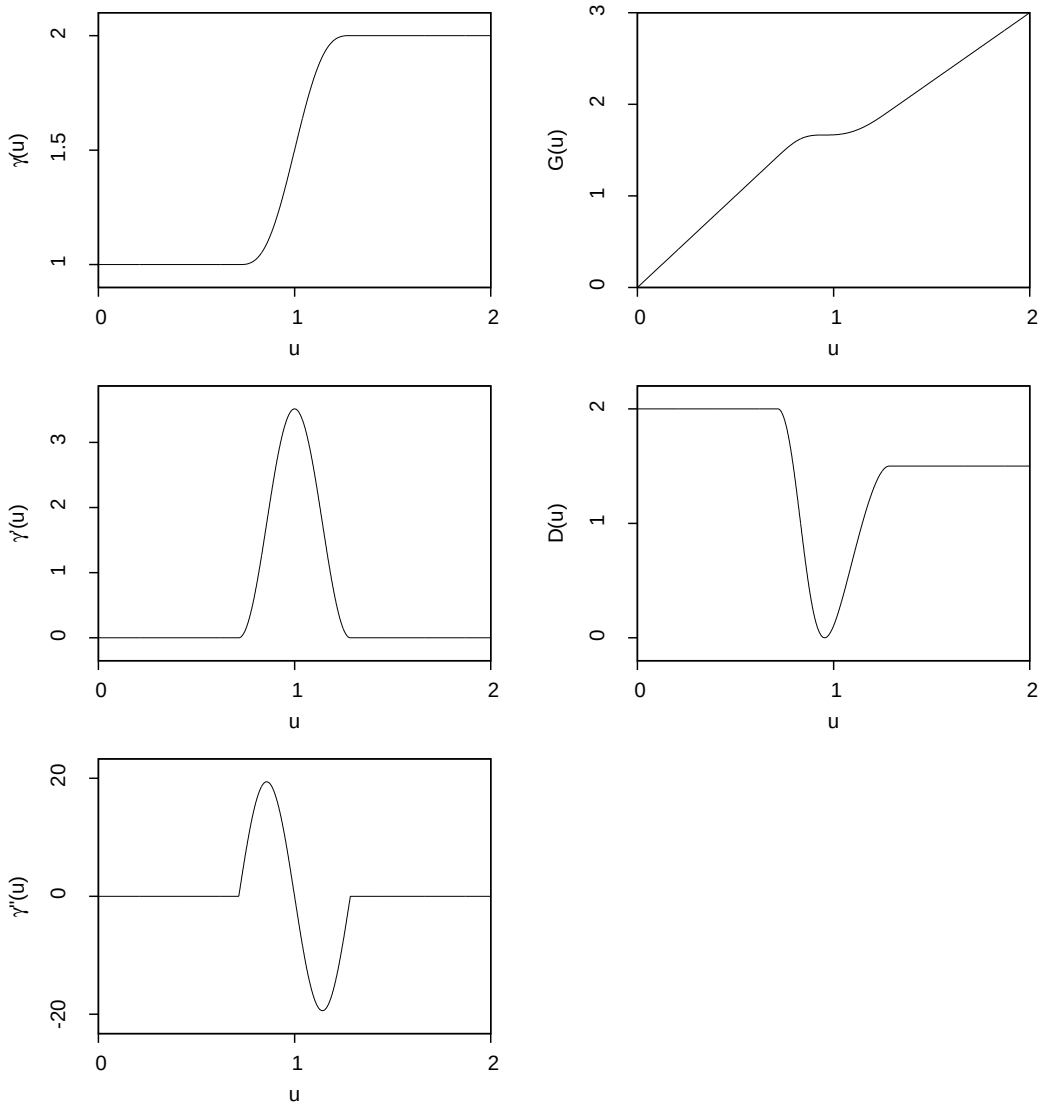

Figure 7.8: Plot of $\gamma(u)$, $\gamma'(u)$, $\gamma''(u)$, $G(u)$ and $D(u)$, Type II, $\Delta_a = 0.568932 = \Delta^*$.

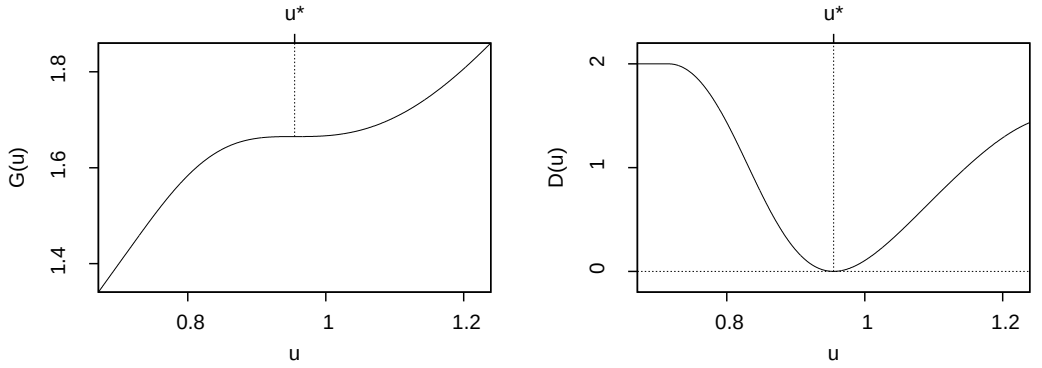

Figure 7.9: Zoom to $G(u)$ and $D(u)$ from Figure 7.8. $\Delta = \Delta^* = 0.568932$, $u^* = 0.954787$.

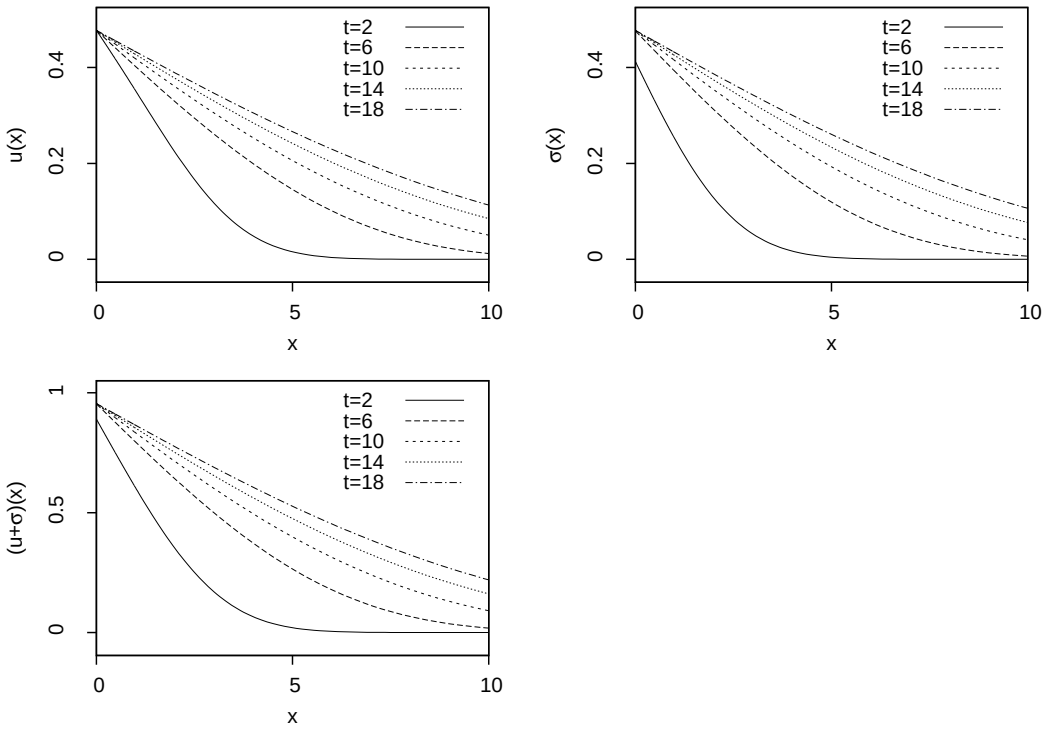

Figure 7.10: Numerical solution $u(x)$, $\sigma(x)$, $(u+\sigma)(x)$ for Type IIa, $\Delta_a = 0.568932 = \Delta^*$. Also $u_s = 0.477394$, hence $u_s < u^*$.

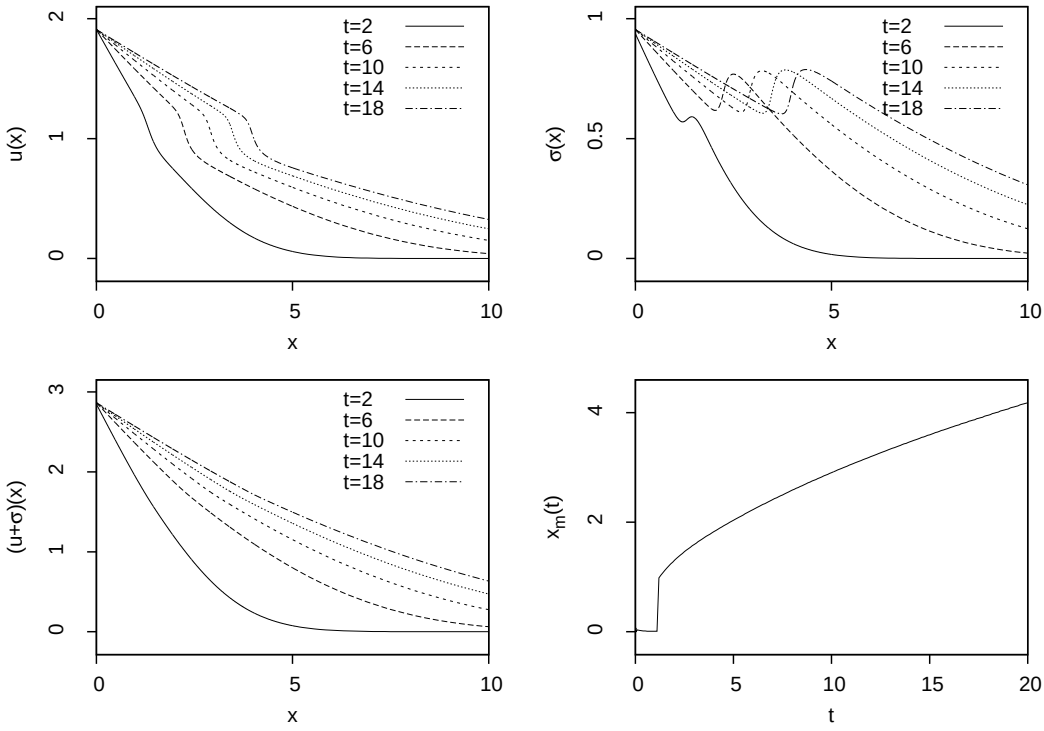

Figure 7.11: Numerical solution $u(x)$, $\sigma(x)$, $(u+\sigma)(x)$ and $x_m(t)$ for Type IId, $\Delta_a = 0.568932 = \Delta^*$. Also $u_s = 1.90957$, hence $u_s > u^*$.

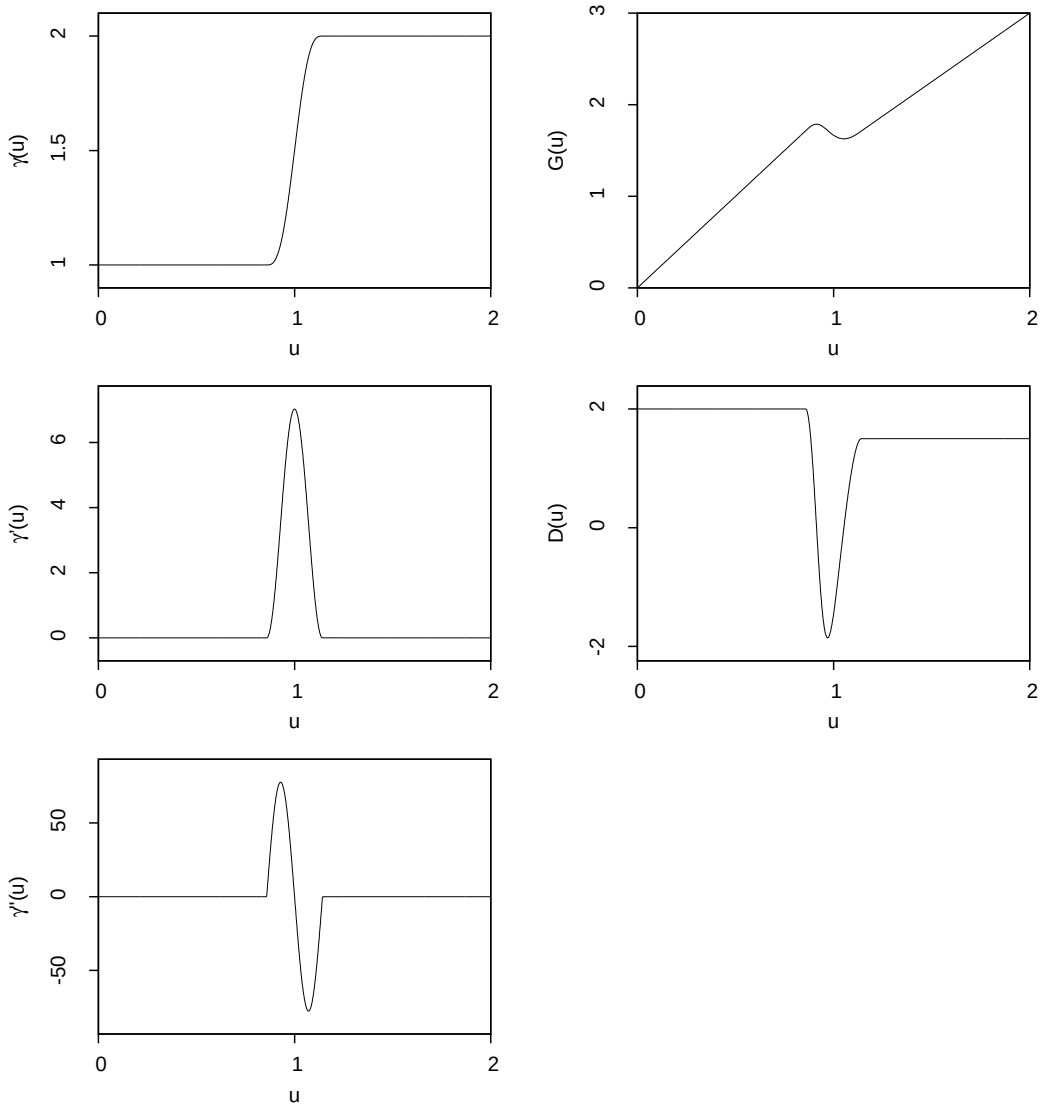

Figure 7.12: Plot of $\gamma(u)$, $\gamma'(u)$, $\gamma''(u)$, $G(u)$ and $D(u)$, Type III, $\Delta_a = 0.284466 < \Delta^*$.

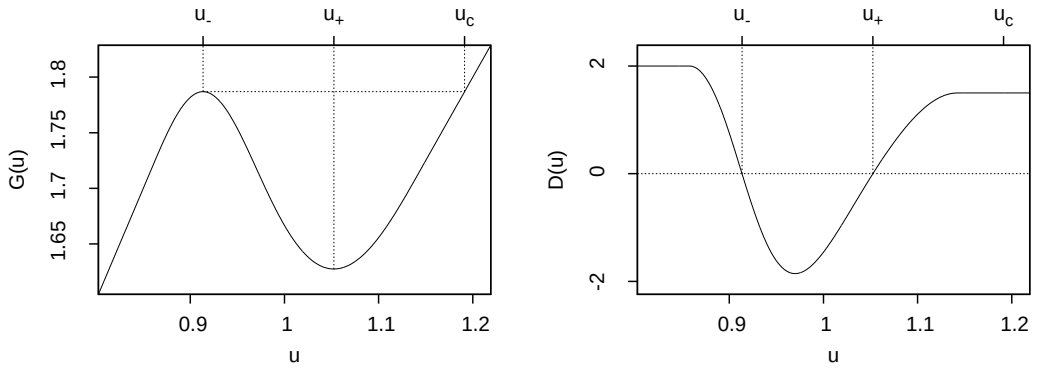

Figure 7.13: Zoom to $G(u)$ and $D(u)$ from Figure 7.12. $\Delta = 0.284466 < \Delta^*$, $u_- = 0.913479$, $u_+ = 1.05248$, $u_c = 1.19129$.

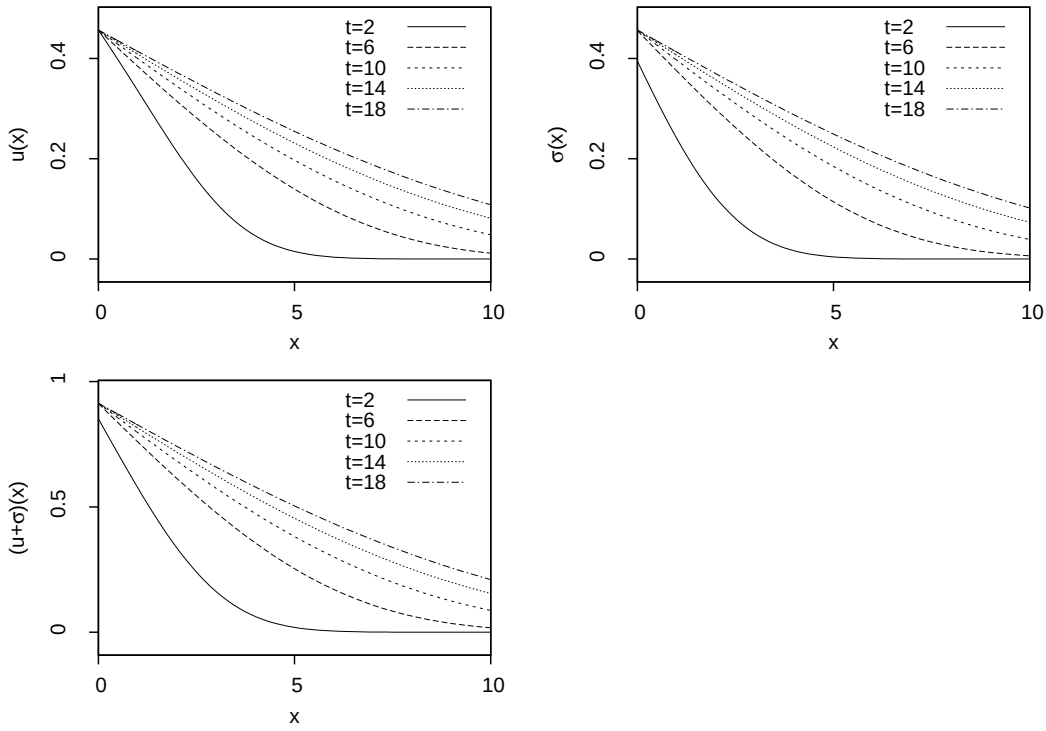

Figure 7.14: Numerical solution $u(x)$, $\sigma(x)$, $(u+\sigma)(x)$ for Type IIIa, $\Delta_a = 0.284466 < \Delta^*$. Also $u_s = 0.456739$, hence $u_s < u_-$.

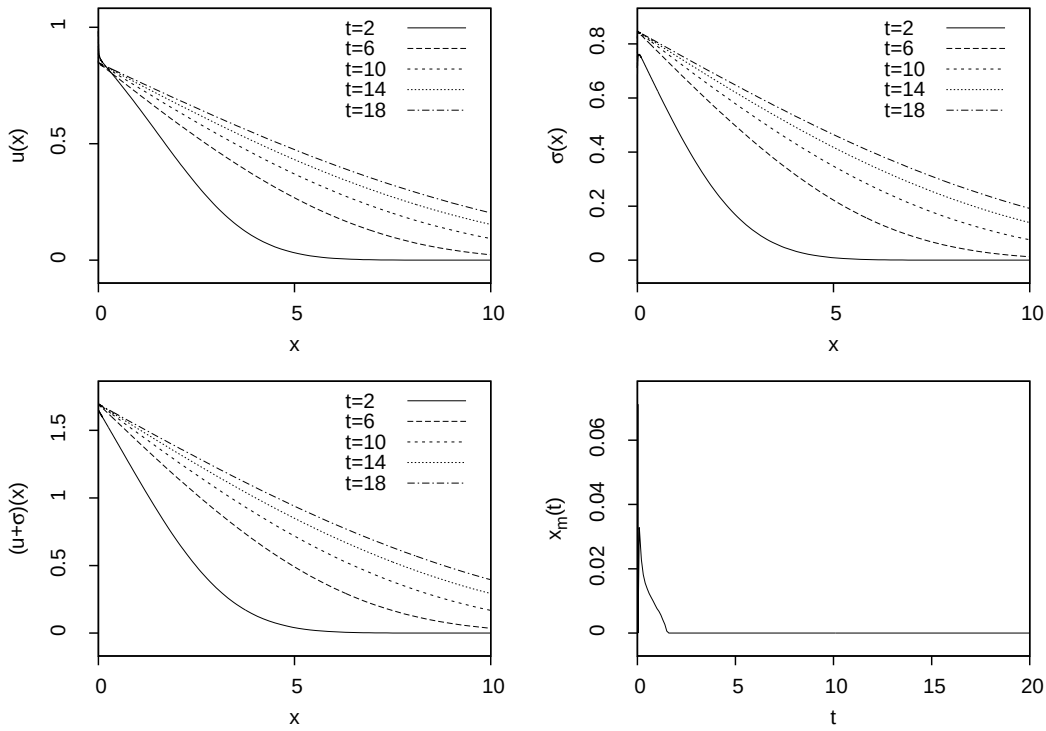

Figure 7.15: Numerical solution $u(x)$, $\sigma(x)$, $(u+\sigma)(x)$ and $x_m(t)$ for Type IIIb, $\Delta_a = 0.284466 < \Delta^*$. Also $u_s = 0.982978$, hence $u_- < u_s < u_+$.

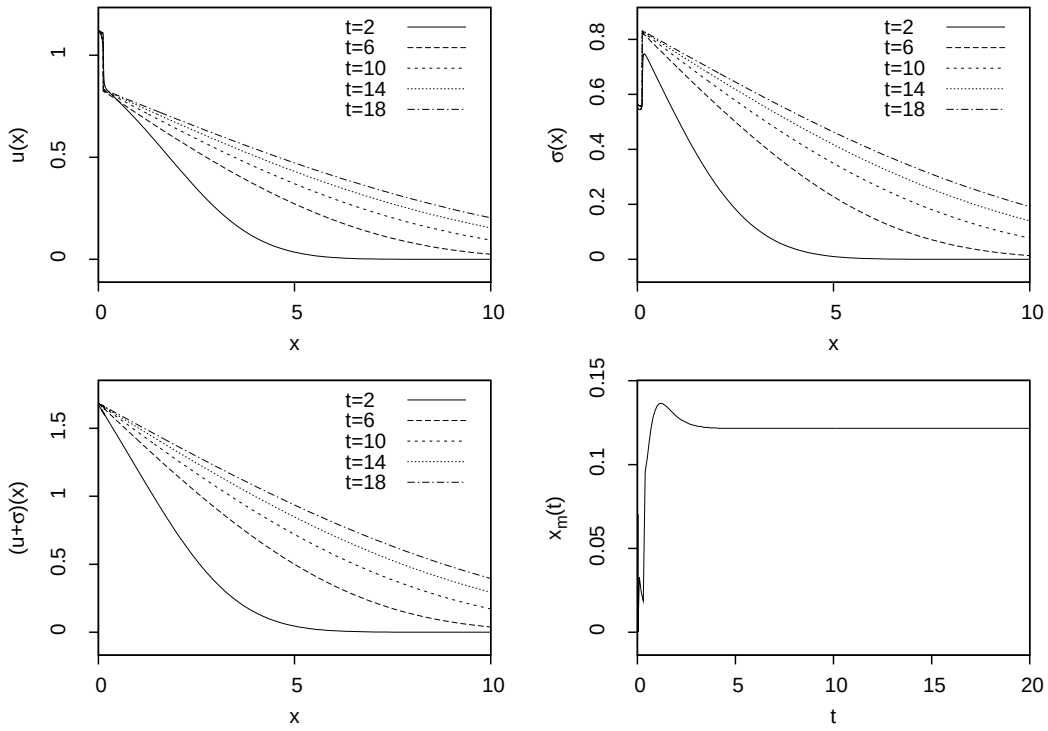

Figure 7.16: Numerical solution $u(x)$, $\sigma(x)$, $(u+\sigma)(x)$ and $x_m(t)$ for Type IIIc, $\Delta_a = 0.284466 < \Delta^*$. Also $u_s = 1.12188$, hence $u_+ < u_s < u_c$.

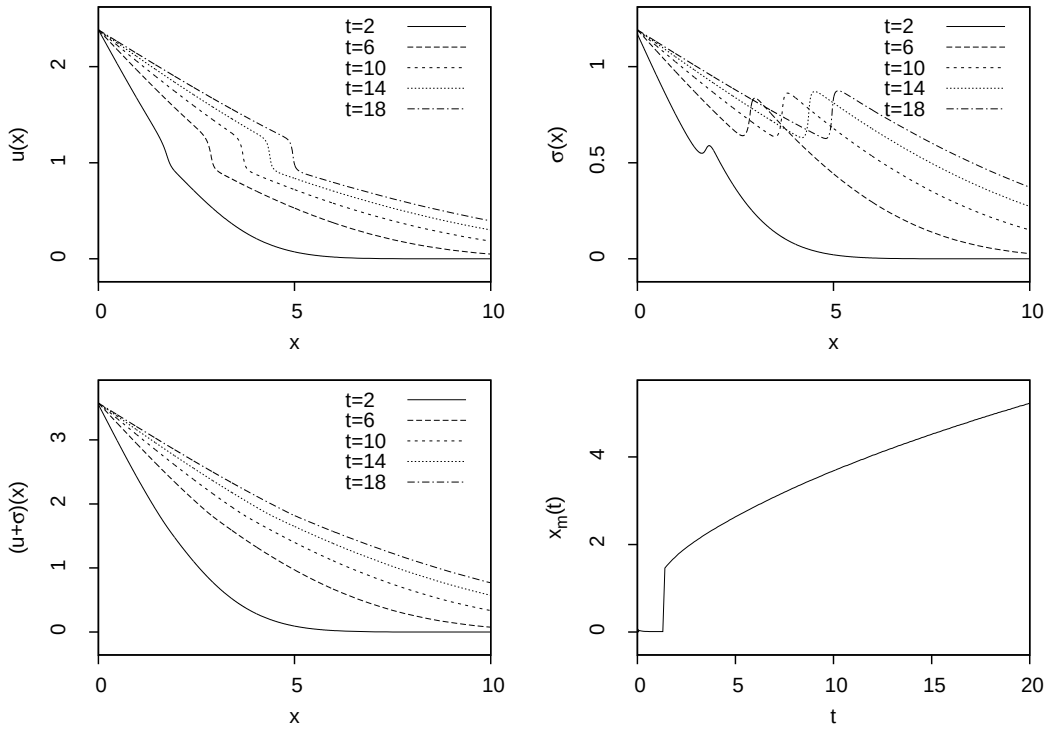

Figure 7.17: Numerical solution $u(x)$, $\sigma(x)$, $(u+\sigma)(x)$ and $x_m(t)$ for Type IIId, $\Delta_a = 0.284466 < \Delta^*$. Also $u_s = 2.38259$, hence $u_c < u_s$.

7.5.3 Wave-like solution in 1D

This subsection is an informal discussion to show that the Cohen model can have some properties similar to the wave equation. This can happen even in the linear case where γ is constant.

We are going to see that the equations of the Cohen model are "close" to the wave equation under certain conditions:

- γ is small,
- the final time T and the domain Ω are large of the same order of magnitude.

More specifically, we assume now that there is a small $\epsilon > 0$ such that $\epsilon T \approx 1$ and that $\epsilon\Omega = \{\epsilon \boldsymbol{x} \ : \ \boldsymbol{x} \in \Omega\}$ has the same order of magnitude as the unit ball in $\mathbb{R}^d$. Also assume that $\gamma(u) = \epsilon\tilde{\gamma}(u)$ for some function $\tilde{\gamma}(u) \approx 1$.

We look at a special case of the Cohen model (3.1a) and (3.1b) where $D = K = \mu = 1$ and $\nu = f = \hat{f} = 0$:

$$\dot{u} - \Delta(u + \sigma) = 0, \tag{7.10a}$$

$$\dot{\sigma} + \gamma(u)\sigma = u. \tag{7.10b}$$

Taking the time derivative of (7.10a) gives

$$\ddot{u} - \Delta(\dot{u} + \dot{\sigma}) = 0.$$

Now rearranging (7.10b) and inserting it gives

$$\ddot{u} - \Delta(\dot{u} + u - \gamma(u)\sigma) = 0. \tag{7.11}$$

We introduce scaled variables $\tilde{t}$ and $\tilde{\boldsymbol{x}}$ that have "normal" magnitude:

$$\tilde{t} = \epsilon\, t, \quad \tilde{\boldsymbol{x}} = \epsilon\, \boldsymbol{x}.$$

With these rescalings, the time derivatives are $\partial/\partial t = \epsilon\, \partial/\partial\tilde{t}$ and $\partial^2/\partial t^2 = \epsilon^2\, \partial^2/\partial\tilde{t}^2$ and the Laplacian becomes $\Delta = \epsilon^2\, \tilde{\Delta}$. Using these in (7.11) and dividing by ϵ^2, we obtain

$$\frac{\partial^2}{\partial\tilde{t}^2}u - \tilde{\Delta}u - \epsilon\tilde{\Delta}\frac{\partial u}{\partial\tilde{t}} - \epsilon\tilde{\Delta}\tilde{\gamma}(u)\sigma = 0.$$

For small values of ϵ, this equation looks like the wave equation

$$\frac{\partial^2}{\partial\tilde{t}^2}u - \tilde{\Delta}u = 0.$$

Note that this derivation of the wave equation is not rigorous. However, it indicates that we can expect wave-like behaviour of the solution under certain conditions.

Example

To demonstrate the wave-like behaviour we give a numerical example. Figure 7.18 shows the numerical solution for small constant γ: $\check{\gamma} = \hat{\gamma} = 0.001$ with $\Omega = (0, 100)$ and $\Gamma_D = \partial\Omega$ where the Dirichlet boundary conditions are $u_D = 1$ at $x = 0$ and $u_D = 0$ at $x = 100$. The final time is chosen as $T = 350$. The discretisation parameters are: $Q = 3$, $r = 3$, $h = 1/256$. To make the movement of the wave front more clearly visible the plot of u and σ is split in two. The first plot shows the time interval $[0, 100]$ when the wave front is advancing and the second plot shows the time interval $[100, 200]$ when the wave is retreating again.

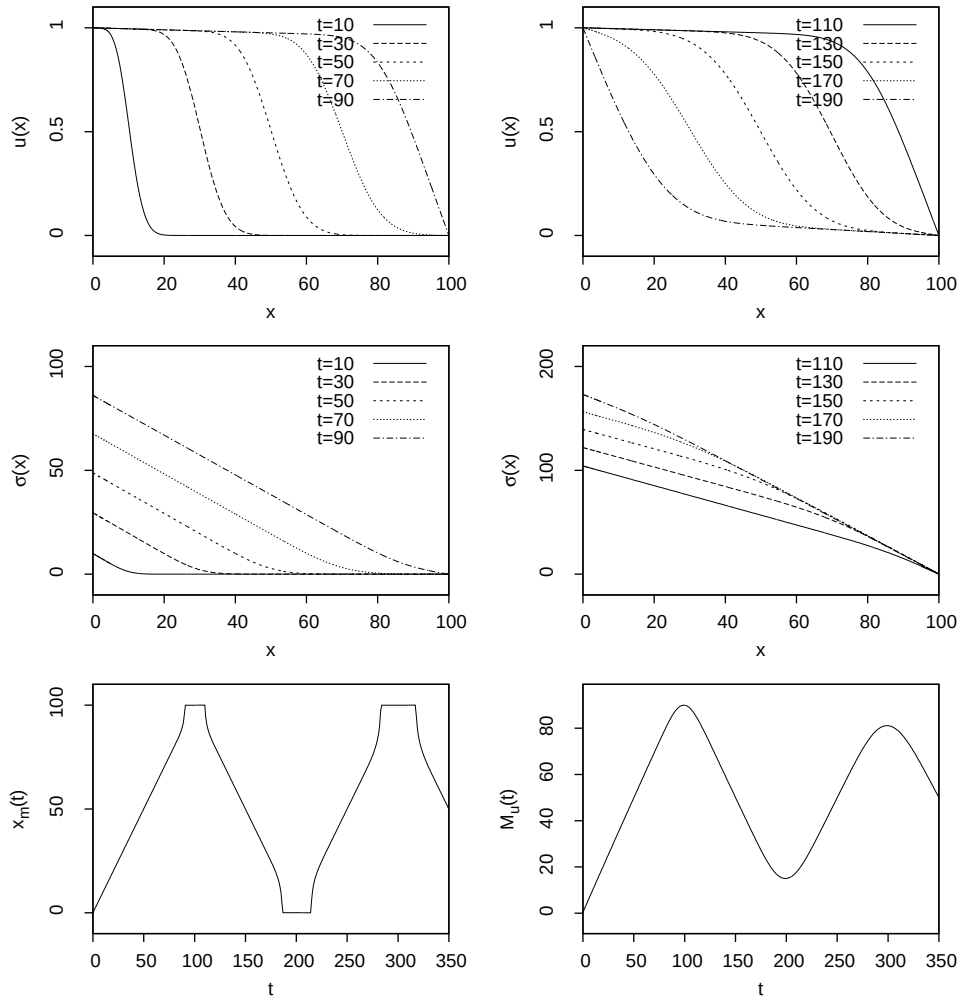

Figure 7.18: Wave-like numerical solution: u and σ for $t \in [0, 100]$ and for $t \in [100, 200]$, position of wave front $x_m(t)$ and mass uptake $M_u(t)$.

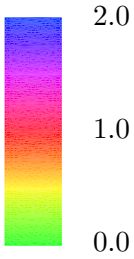

Figure 7.19: Colour map for the following plots, the function value is indicated by the colour.

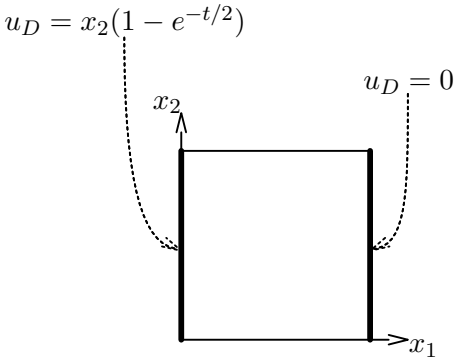

Figure 7.20: Domain for 2D example, $\Omega = (0,1) \times (0,1)$, Dirichlet boundary conditions on opposite sides of the square.

7.5.4 Sharp front in 2D

Similar to Subsection 7.5.1, we consider a two-dimensional example where the solution develops a sharp front. The domain is $\Omega = (0,1)^2$ with $\Gamma_D = \{0,1\} \times (0,1) = \{(x_1, x_2) : (x_1 = 0 \vee x_1 = 1) \wedge x_2 \in (0,1)\}$, see Figure 7.20. The Neumann boundary condition is homogeneous: $g = 0$, whereas the Dirichlet boundary conditions are: $u_D = x_2(1-e^{-t/2})$ at $x_1 = 0$ and $u_D = 0$ at $x_1 = 1$. The other parameters of the problem are $\gamma = \gamma_3$, $\check{\gamma} = 0.5$, $\hat{\gamma} = 1.0$, $u_a = 0.5$, $\Delta_a = 0.05$ and $T = 30$. The discretisation parameters are $Q = 3$, $k = 0.015$, $h = 1/32$ and $r = 2$. Figures 7.21 and 7.22 show the numerical solution for u at $t \in \{1.8, 3.6, 15, 30\}$. Figures 7.23 and 7.24 show the corresponding plots for σ, and Figures 7.25 and 7.26 those for $u + \sigma$. The function values are indicated both by colour and by the z-value of the plot. Figure 7.19 shows which colour corresponds to which number. As in the one-dimensional example u develops a sharp front, which remains stationary.

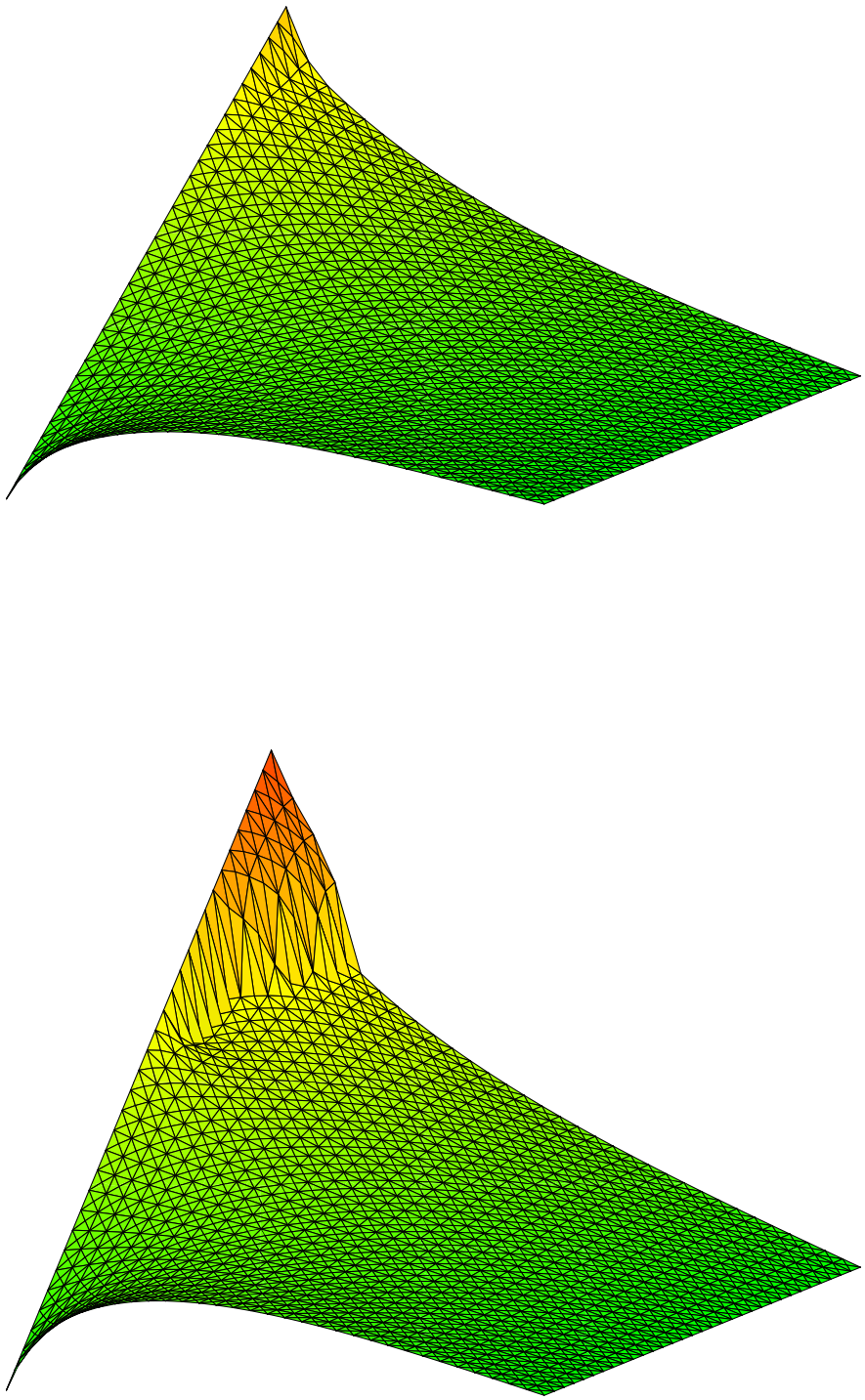

Figure 7.21: Sharp front in 2D: Numerical solution u at $t = 1.8$ (top) and $t = 3.6$ (bottom).

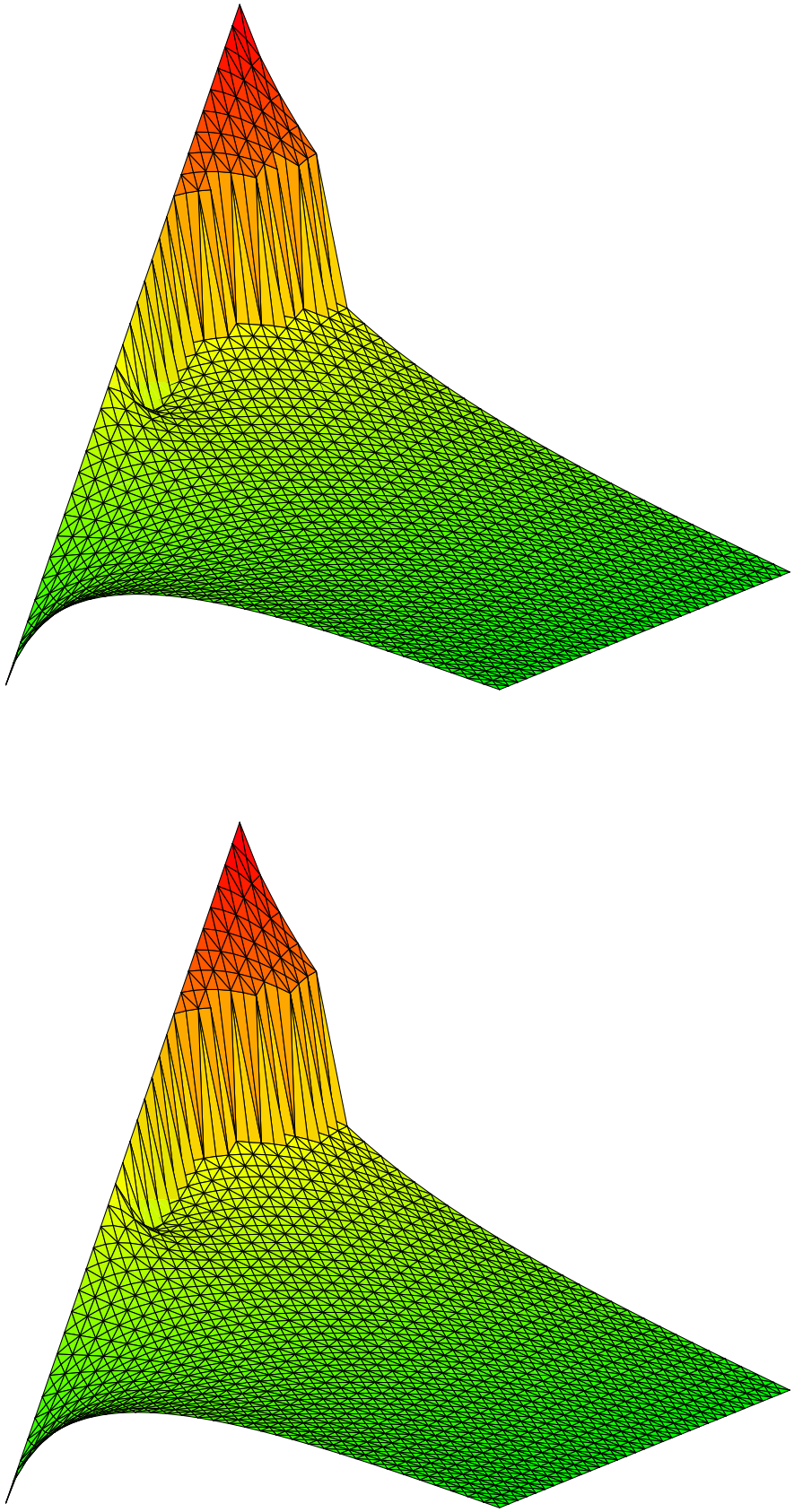

Figure 7.22: Sharp front in 2D: Numerical solution u at $t = 15$ (top) and $t = 30$ (bottom).

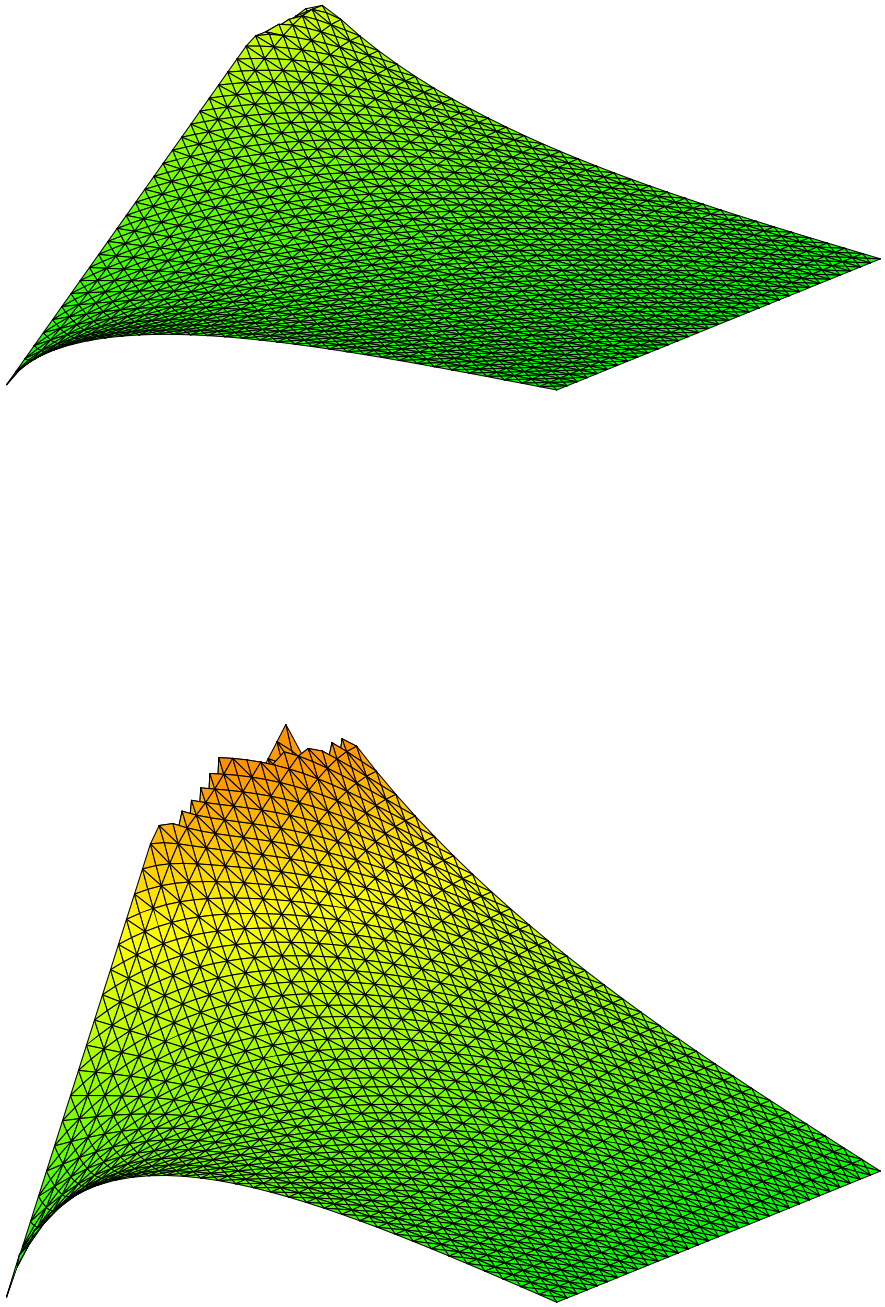

Figure 7.23: Sharp front in 2D: Numerical solution σ at $t = 1.8$ (top) and $t = 3.6$ (bottom).

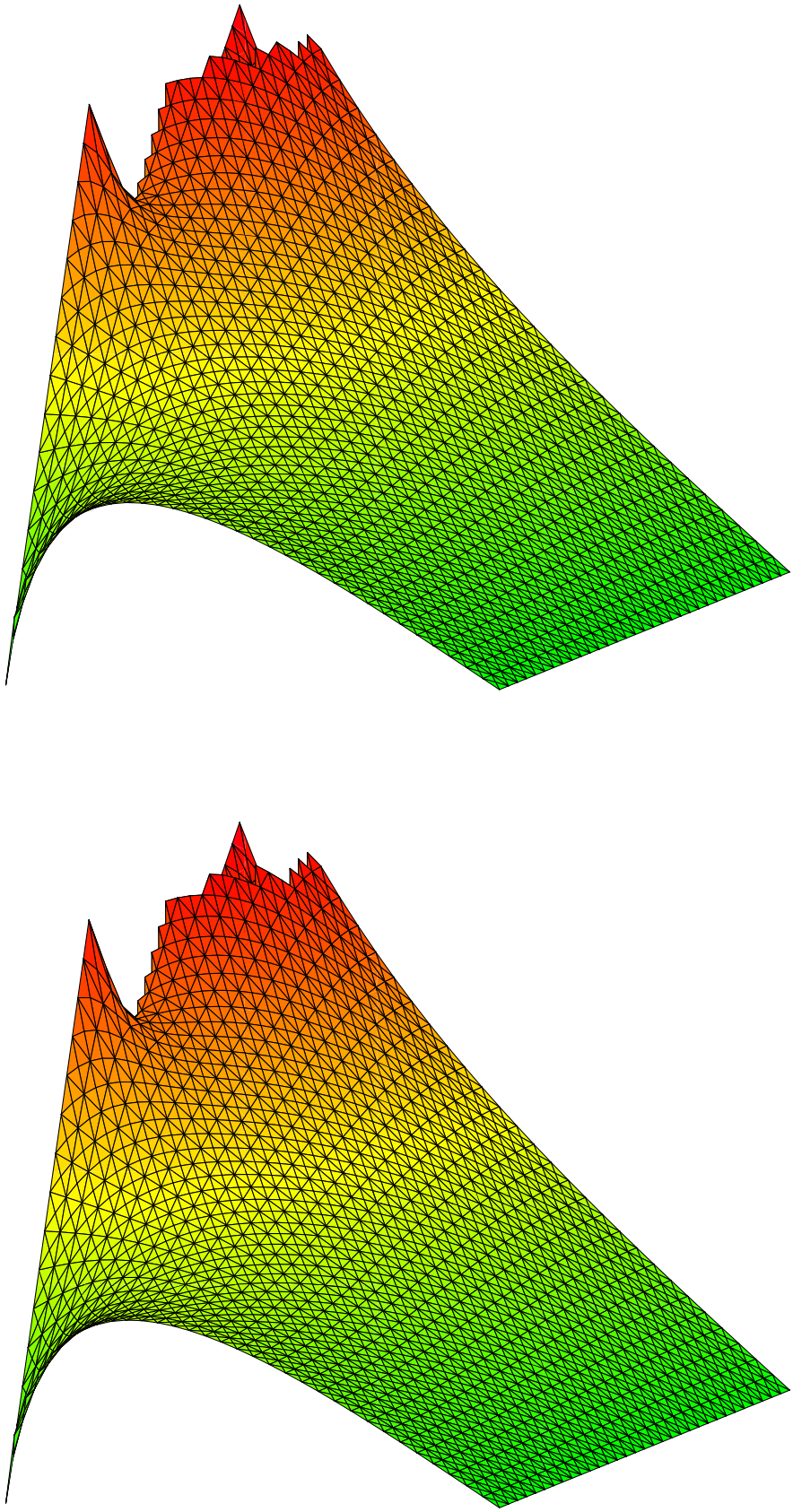

Figure 7.24: Sharp front in 2D: Numerical solution σ at $t = 15$ (top) and $t = 30$ (bottom).

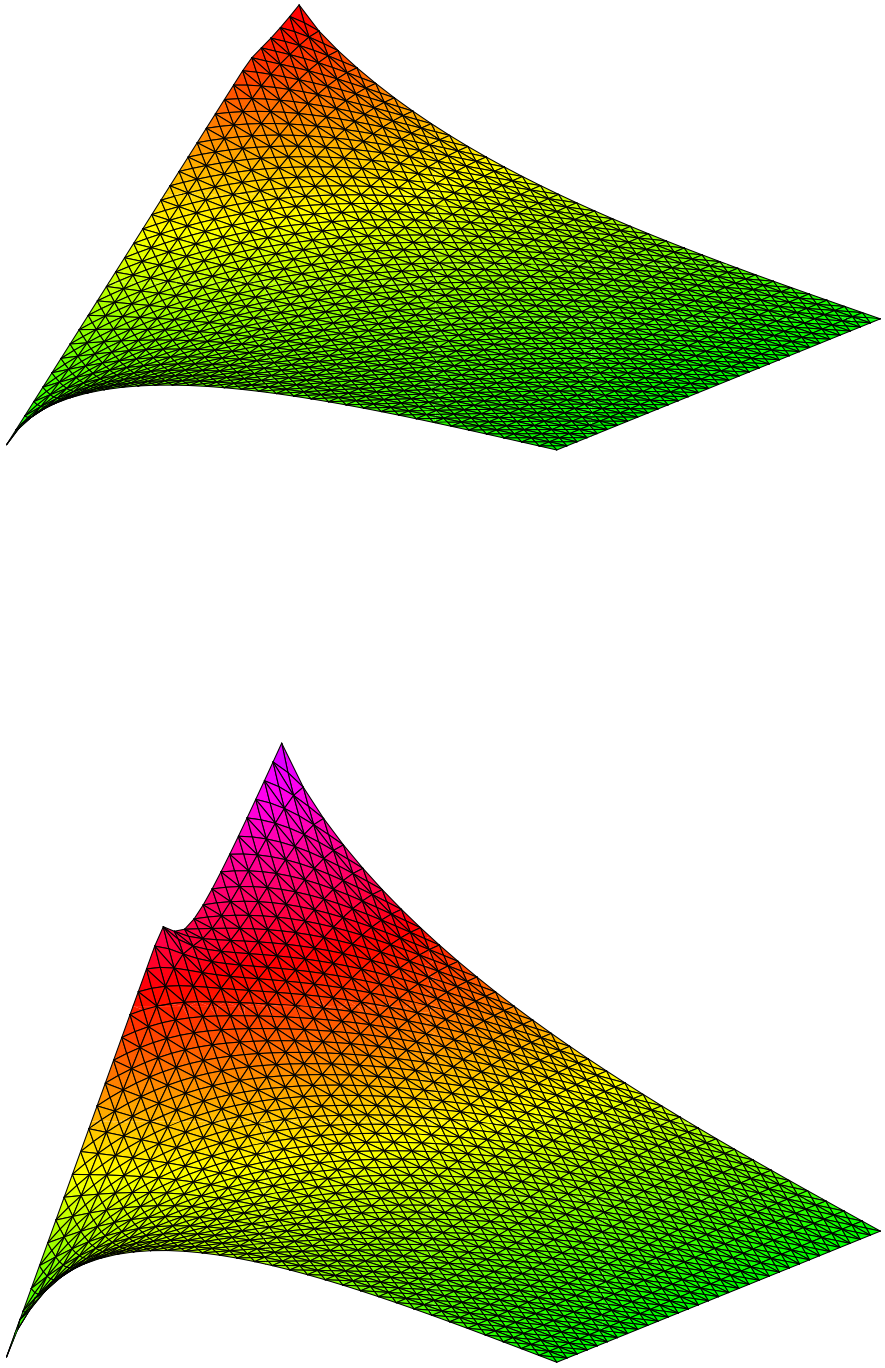

Figure 7.25: Sharp front in 2D: Numerical solution $u + \sigma$ at $t = 1.8$ (top) and $t = 3.6$ (bottom).

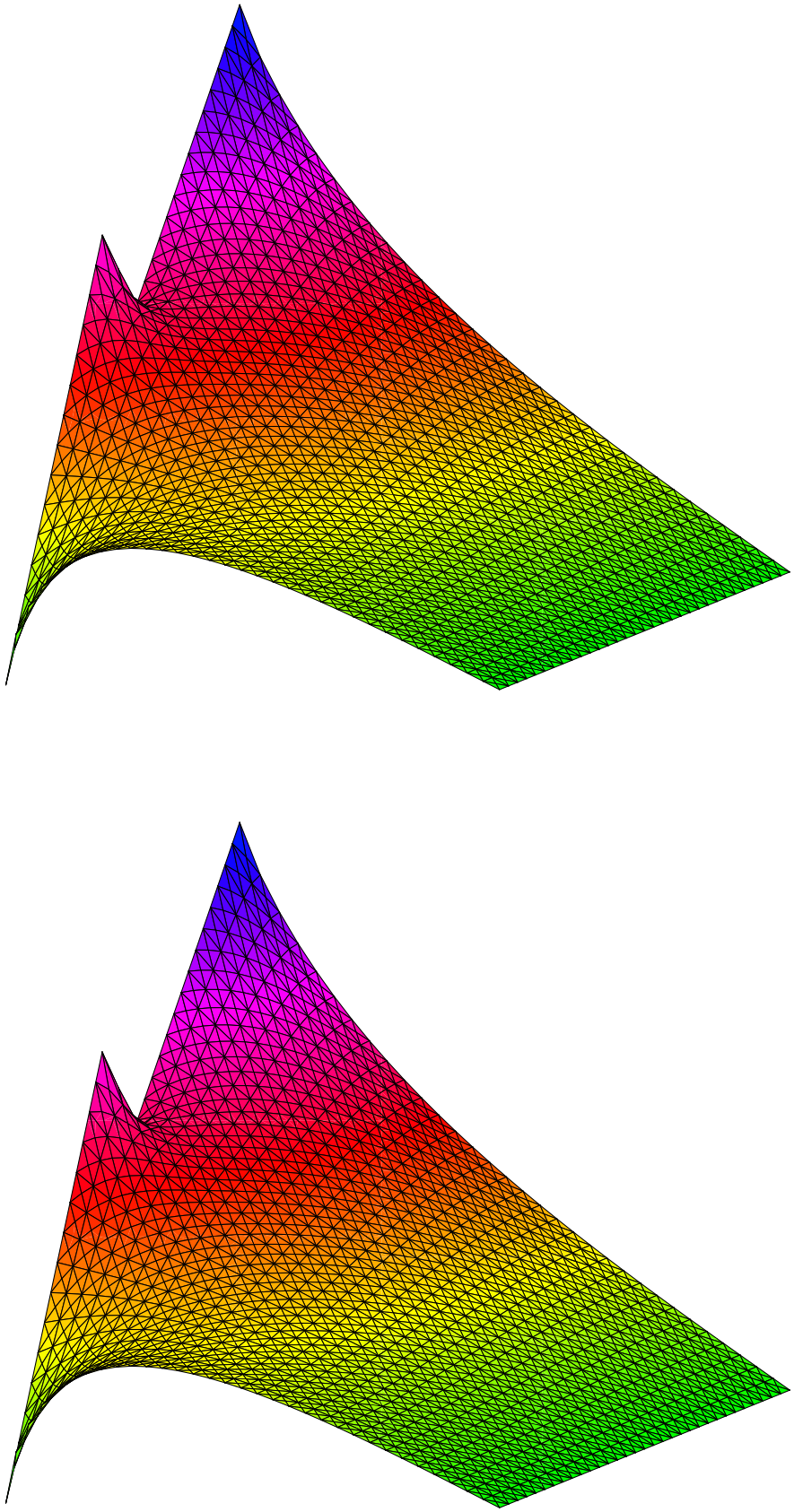

Figure 7.26: Sharp front in 2D: Numerical solution $u + \sigma$ at $t = 15$ (top) and $t = 30$ (bottom).

7.5.5 Ring-shaped domain in 2D

In this subsection we present another numerical experiment in 2D, on the ring-shaped domain $\Omega = \{\boldsymbol{x} \in \mathbb{R}^2 \ : \ 1 < |\boldsymbol{x}| < 3\}$. The boundary conditions are Dirichlet only with $u_D = 0$ for $|\boldsymbol{x}| = 3$ and $u_D(\boldsymbol{x}) = 1 + x_2$ for $|\boldsymbol{x}| = 1$, see Figure 7.27. The other parameters are: $\gamma = \gamma_3$, $\check{\gamma} = 0.5$, $\hat{\gamma} = 3$, $u_a = 0.5$, $\Delta = 0.05$. The discretisation parameters are: $Q = 3$, $r = 3$, $k = 0.01$ and $h = 0.19$. See Figures 7.28 and 7.29 for a plot of u, σ and $u + \sigma$ at $t = 20$. At this time, the sharp front in u has developed and the solution has become stationary. Note that, as in the other examples, $u + \sigma$ does not have a sharp front.

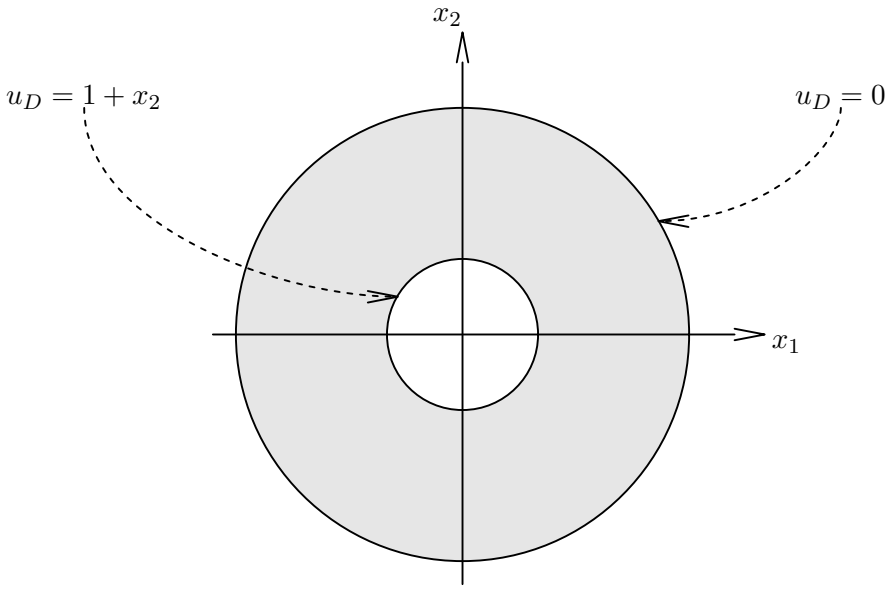

Figure 7.27: Ring-shaped domain in 2D, $\{\boldsymbol{x} \in \mathbb{R}^2 \ : \ 1 < |\boldsymbol{x}| < 3\}$, with Dirichlet boundary conditions.

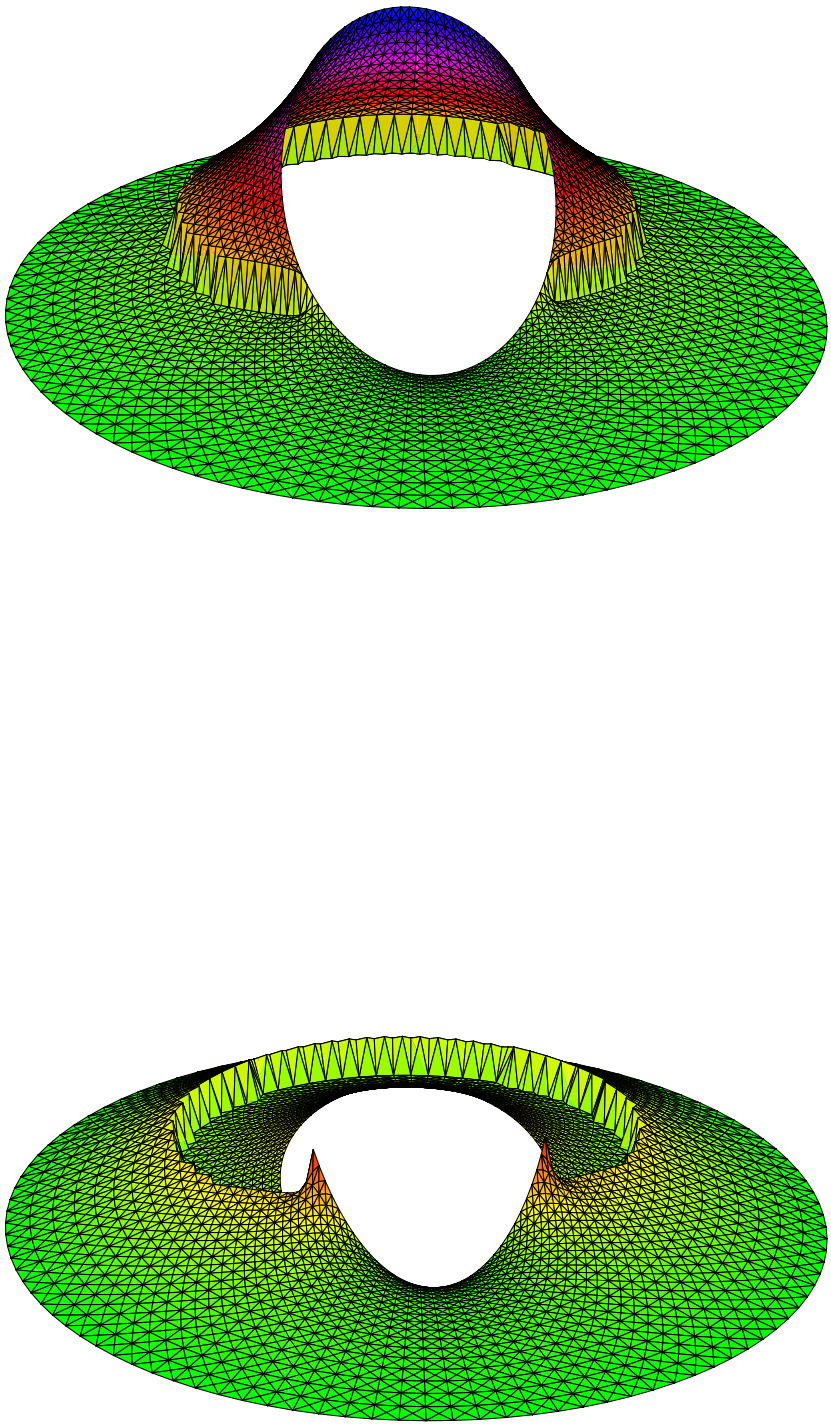

Figure 7.28: Ring-shaped domain in 2D: Numerical solution u (top) and σ (bottom) at $t = 20$.

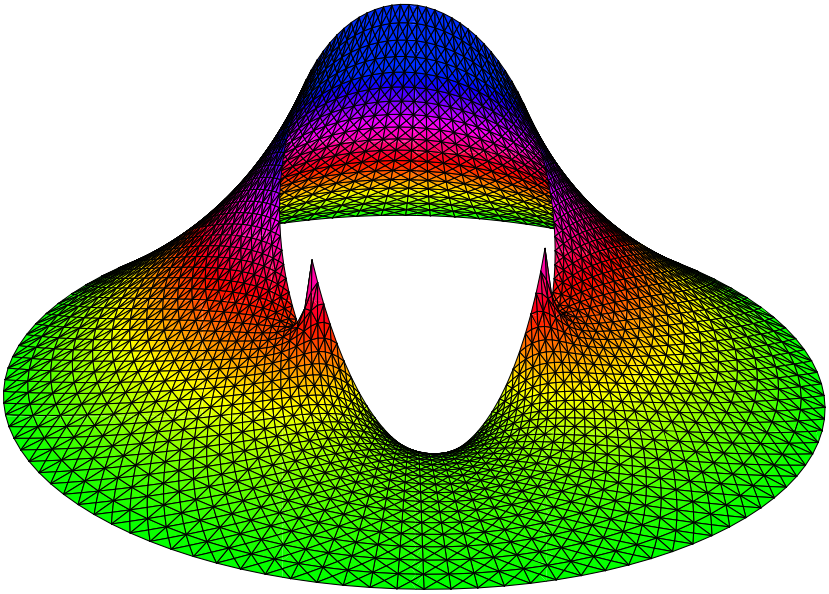

Figure 7.29: Ring-shaped domain in 2D: Numerical solution $u + \sigma$ at $t = 20$.

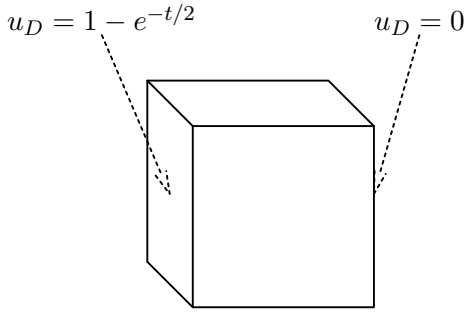

Figure 7.30: Domain for 3D example, $\Omega = (0,1)^3$, Dirichlet boundary conditions on opposite faces of the cube.

7.5.6 Sharp front in 3D

As the final example, we consider the three-dimensional equivalent to the example from Subsection 7.5.1. The domain is $\Omega = (0,1)^3$ with $\Gamma_D = \{0,1\} \times (0,1)^2$, see Figure 7.30. The Neumann boundary condition is homogeneous: $g = 0$, whereas the Dirichlet boundary conditions are: $u_D = 1$ at $x_1 = 0$ and $u_D = 0$ at $x_1 = 1$. The other parameters of the problem are $\gamma = \gamma_3$, $\check{\gamma} = 0.5$, $\hat{\gamma} = 1.0$, $u_a = 0.5$, $\Delta_a = 0.05$ and $T = 10$. The discretisation parameters are $Q = 3$, $k = 0.01$, $h = 1/16$ and $r = 2$. As in the other examples on a "small" domain u develops a stationary sharp front. See Figure 7.31 for the u and σ at $t = 10$ and Figure 7.32 for $u + \sigma$ at $t = 10$. In these plots the function values are indicated by colour only.

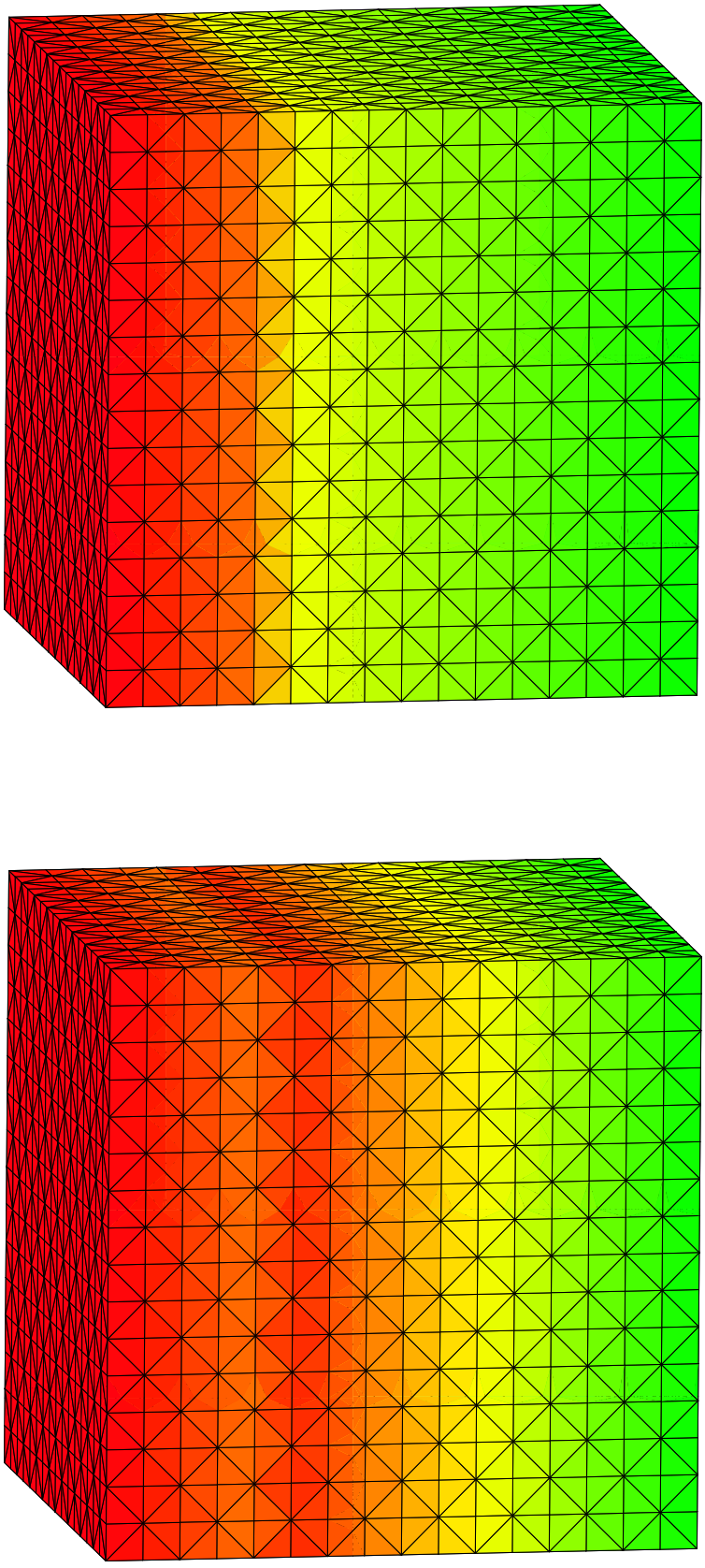

Figure 7.31: Sharp front in 3D: Numerical solution u (top) and σ (bottom) at $t = 10$.

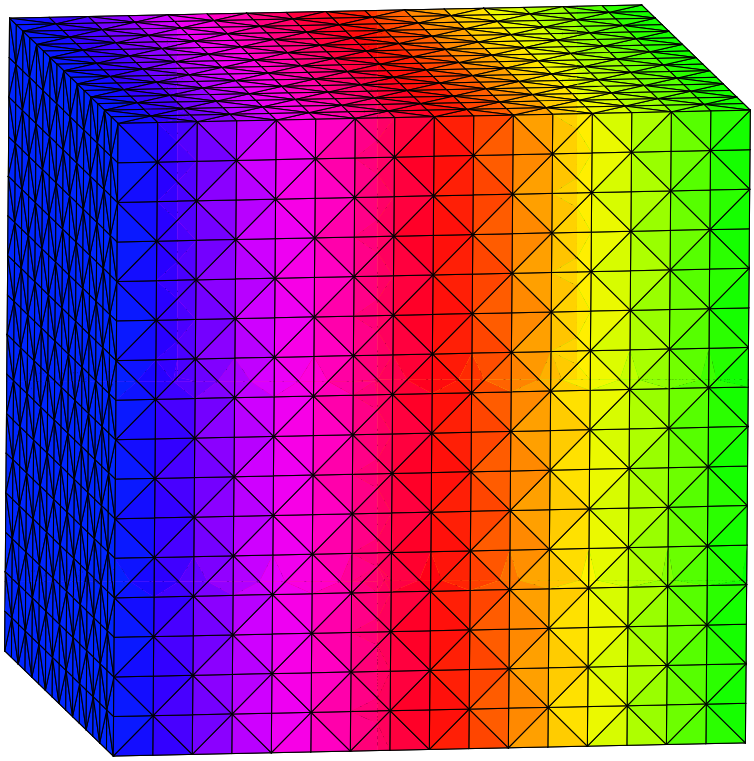

Figure 7.32: Sharp front in 3D: Numerical solution $u + \sigma$ at $t = 10$.

Chapter 8

Conclusion

In this chapter we review the results of this thesis and make suggestions for further work.

In Chapter 4 we simplified the standard variational formulation and prescribed homogeneous Dirichlet boundary conditions for σ. This way we could take a first step towards proving well-posedness of the problem: the linearised problem has a unique solution and the nonlinear problem in 1D has a unique solution.

In Chapter 5 we used the simplified variational formulation as a basis for a discrete formulation. Five methods for discretising the nonlinearity were considered, two of which were nonlinear ($Q = 4, 5$), the other three used extrapolation from previous time steps.

For four of the five methods boundedness results were proved in Chapter 5 and an a priori estimate (Theorem 6.8) was proved in Chapter 6, which is the central result of this thesis. In order to prove this a priori estimate the two approaches introduced in Subsection 4.1.1 were combined: the use of $u + \sigma$ provided norms which were strong enough to make all the steps of the proof possible, while the inverse Laplacian removed the 'cross terms' involving u and σ, which effectively removed a maximum bound on the time step size for the linearised methods, see Subsection 6.4.1.

For the remaining method $Q = 3$, no theoretical results could be given in contrast to the numerical results in Chapter 7, where $Q = 3$ appears to have the same convergence rates as the other methods.

The nonlinear methods appeared problematic from both the theoretical point of view as well as from the numerical evidence: first, we saw that the discrete solution may be non-unique (Section 5.4). Then, the a priori error estimate required the time step to be below a certain constant $\hat{k}$. Furthermore, the numerical results of Chapter 7 confirm this difficulty. It appears as if exactness and h-convergence only work for small enough time steps.

8.1 Suggestions for further work

We now make some remarks about how the work of this thesis can be extended. We also include open questions which arose during the work on this thesis.

- As mentioned before, theoretical results for method $Q = 3$ that uses a product approximation are missing. The difficulties and possible ways forward have been discussed in Sections 5.4 and 6.5. Since this method seems to work so well from the numerical point of view, it would be good to be able to confirm the convergence rates with theoretical results.

- For the case where $\nu \neq 0$ in (3.1b), i.e. where the equation for σ contains the extra term $\nu\dot{u}$ on the right-hand side, we have neither theoretical nor numerical results. As many of the articles on the Cohen model, for example [Edw95], [HZ96], retain this term, it seems to be an interesting generalisation.

- Both from the theory and the numerics the combined variable $y := u + \sigma$ appears to be smoother than either u or σ on their own. This suggests that it might be useful to rewrite the model in terms of y and σ and discretise these two variables instead of u and σ. See also the Proof of Proposition 4.10.

- Because of using the ALBERTA library the numerical calculations are done using double precision floating point numbers (64 bit). For some of the convergence tables, for example Table 7.94, it would be interesting to see if higher accuracy can be achieved by using floating point numbers of higher precision.

- A question that can not be answered with the tools used in this thesis but which appeared when studying the numerical experiments is:

  Are there choices of parameters such that on a small domain the sharp front that we saw in the numerical examples does not become stationary but continues to move?

  We saw that on a large domain the sharp front can continue to move, but with decreasing speed, see for example Figure 7.17 in Subsection 7.5.2. Also, in Figure 7.18 in Subsection 7.5.3 we have seen an example of a wave-like solution, where the wave front moved at constant speed. In this example, however the wave front is not steep, the steepness is even decreasing with time.

  Thus, in all examples with a sharp front on a small domain, that were presented in this thesis, the sharp front became stationary. The question is whether this is necessarily so or whether the sharp front can also move with constant speed. See also the articles [Edw95] and [SE04], which look at this question on an unbounded domain.

Appendix A

More on Mathematical Background

This chapter of the appendix is a collection of some mathematical results. They are collected here in the appendix since they either do not really fit into Chapter 2, which gives the background from functional analysis and finite element theory, or since they are more detailed than necessary for that chapter.

A.1 Basic inequalities

In this thesis we only need the following special case of Young's inequality, which can be derived directly from $(\sqrt{\epsilon}a - \frac{1}{\sqrt{\epsilon}}b)^2 \geq 0$.

For $a, b \in \mathbb{R}$ and any $\epsilon > 0$,

$$2ab \leq \epsilon\, a^2 + \frac{1}{\epsilon}\, b^2. \tag{A.1}$$

For convenience we state a second form of Young's inequality, which is of course equivalent to the first one:

For $a, b \in \mathbb{R}$ and any $\epsilon > 0$,

$$ab \leq \epsilon\, a^2 + \frac{1}{4\epsilon}\, b^2. \tag{A.2}$$

Another inequality which is needed in this thesis is Jensen's inequality. Here we state it in a form as a special case of [Zhe04, (1.3.42)]. For a definition of convex functions, see Section A.3.

Lemma A.1 (Jensen's inequality).
Let $f : \mathbb{R} \to \mathbb{R}$ be a convex function and $y : [a, b] \to \mathbb{R}$. Then

$$f\left(\frac{1}{b-a}\int_a^b y(x)\,dx\right) \leq \frac{1}{b-a}\int_a^b f\left(y(x)\right)\,dx. \tag{A.3}$$

The following versions of the continuous and discrete Grönwall lemma are taken from [SWW04, Section 2.6]:

Lemma A.2 (Continuous Grönwall lemma).
Let u be a non-negative integrable function defined on $\mathcal{J} := [0, T]$ and v be a non-negative and non-decreasing function also defined on $\mathcal{J}$. Then the inequality

$$u(t) \leq v(t) + C\int_0^t u(s)\,ds \quad \forall\, t \in \mathcal{J},$$

where $C \geq 0$ is a constant, implies that $u(t) \leq v(t)\exp(Ct)$ for all $t \in \mathcal{J}$.

For convenience we state a direct consequence of the Grönwall lemma. It is this form which is needed throughout the thesis.

Lemma A.3 (Continuous Grönwall lemma, second form).
If, for nonnegative integrable functions $\phi, \psi, \theta : [0,T] \to \mathbb{R}$ with θ nondecreasing and for $\alpha \geq 0$,

$$\phi(t) + \psi(t) \leq \theta(t) + \alpha \int_0^t \phi(s)\, ds,$$

then for $t \in [0,T]$ we have

$$\phi(t) + \psi(t) \leq \exp(\alpha T)\, \theta(t).$$

Proof. Since $\psi(t) \geq 0$, we can apply Lemma A.2 to $u := \phi + \psi$, noting that $\exp(\alpha t) \leq \exp(\alpha T)$. □

We also need a discrete version of Grönwall's lemma:

Lemma A.4 (Discrete Grönwall lemma).
For a positive integer N let u be a non-negative function defined on $\mathcal{J}^k := \{t_i \in \mathcal{J} : t_i = ik, i = 0, \ldots, N = T/k\}$ with $u_i := u(t_i)$ and v be a non-negative and non-decreasing function also defined on $\mathcal{J}^k$. Then the inequality

$$u_i \leq v_i + Ck \sum_{j=0}^{i-1} u_j \qquad \forall t_i \in \mathcal{J}^k,$$

where $C \geq 0$ is a constant, implies that $u_i \leq v_i \exp(Ct_i)$ for all $t_i \in \mathcal{J}^k$.

Proof. We use induction. For $i = 0$, the assertion is obvious. Now, assume that $u_j \leq v_j \exp(Ct_j)$ for all $j = 1, \ldots, i$, and use this to prove $u_{i+1} \leq v_{i+1} \exp(Ct_{i+1})$:

$$\begin{aligned} u_{i+1} &\leq v_{i+1} + Ck \sum_{j=0}^{i} u_j \leq v_{i+1} + Ck \sum_{j=0}^{i} v_j \exp(Ct_j) \\ &\leq v_{i+1}(1 + Ck \sum_{j=0}^{i} \exp(Ct_j)) \end{aligned}$$

(since v_j is non-decreasing),

$$= v_{i+1}(1 + Ck \sum_{j=0}^{i} (\exp(Ck))^j) \leq v_{i+1}(1 + (\exp(Ck) - 1) \sum_{j=0}^{i} (\exp(Ck))^j)$$

(since $x \leq \exp(x) - 1$ for all $x \in \mathbb{R}$),

$$= v_{i+1}(1 + ((\exp(Ck))^{i+1} - 1)) = v_{i+1}(\exp(Ct_{i+1}).$$ □

A.2 Banach's fixed point theorem

The standard version of Banach's fixed point theorem is formulated for complete metric spaces. But for this thesis we only need the theorem for closed subsets of Banach spaces, which is an important class of complete metric spaces. Therefore we state the fixed point theorem for this special case:

Theorem A.5 (Banach's fixed point theorem).
Let $(Y, \|\cdot\|_Y)$ be a Banach space, and $Z \subseteq Y$ a closed subset. If $A : Z \to Z$ is a contraction, i.e. a mapping for which there is $0 \leq q < 1$ such that

$$\|A(x) - A(y)\|_Y \leq q \cdot \|x - y\|_Y \quad \forall x, y \in Z,$$

then A has exactly one fixed point $z \in Z$.

Proof. For the more general version of the theorem in complete metric spaces, see [OD96, Theorem 4.10.1], for example. A statement like here, for Banach spaces, can be found in [BG87b, Theorem 1.31]. □

A.3 Bounds for norms in finite-dimensional spaces

This section has the aim of proving sharp constants of the equivalence of the p-norms in $\mathbb{R}^n$, Lemma A.14.

Definition A.6 (Convex and concave functions).
Let X be a vector space, $K \subseteq X$ a convex subset and $f : K \to \mathbb{R}$ a function. f is called *convex* iff

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y) \tag{A.4}$$

for all $\theta \in [0,1]$ and $x, y \in K$. f is called *concave* iff

$$f(\theta x + (1-\theta)y) \geq \theta f(x) + (1-\theta) f(y) \tag{A.5}$$

for all $\theta \in [0,1]$ and $x, y \in K$.

Remark A.7.
Note that the case of equality is included in both definitions. This implies that a linear function $f : K \to \mathbb{R}$ is both convex and concave.

Calculus gives an easy criterion for testing convexity and concavity of functions. Here, we cite a result from [Heu01, Satz 49.8] for the one-dimensional case:

Proposition A.8 (Criterion for convexity/concavity).
Let $f : \mathbb{R} \to \mathbb{R}$ be continuous on the interval I and twice differentiable in its interior $\mathring{I}$.
If $f''(x) \geq 0$ on $\mathring{I}$, then f is convex on I.
If $f''(x) \leq 0$ on $\mathring{I}$, then f is concave on I.

Corollary A.9 (x^p is convex if $p \geq 1$).
Let $\alpha > 0$ and let

$$f : \left\{ \begin{array}{rcl} \mathbb{R}_{\geq 0} & \to & \mathbb{R} \\ x & \mapsto & x^\alpha \end{array} \right.$$

If $\alpha \geq 1$, then f is convex. If $0 < \alpha \leq 1$, then f is concave.

Proof. Apply Proposition A.8, with $f''(x) = \alpha(\alpha - 1)x^{\alpha - 2}$ for $\alpha \neq 1$. □

The next proposition is taken from an exercise in Bourbaki [Bou04, Chapter 1, Ex. §4, 25b].

Proposition A.10 (Sub- and superadditivity).
Let $f : \mathbb{R}_{\geq 0} \to \mathbb{R}$ be a function with $f(0) = 0$. If f is convex, then it is superadditive, *i.e.*

$$f(x + y) \geq f(x) + f(y) \quad \forall\, x, y \in \mathbb{R}_{\geq 0}.$$

If f is concave then then it is subadditive, *i.e.*

$$f(x + y) \leq f(x) + f(y) \quad \forall\, x, y \in \mathbb{R}_{\geq 0}.$$

Proof. First let f be convex. Choosing $y = 0$ in (A.4) gives

$$f(\theta x) \leq \theta f(x) \quad \forall\, \theta \in [0,1]\ \forall\, x \in \mathbb{R}_{\geq 0}. \tag{A.6}$$

Now let $x \in \mathbb{R}_{\geq 0}$ and $\lambda \geq 1$. Then, since $\lambda^{-1} \in (0,1]$, using (A.6),

$$f(\lambda x) = \lambda\lambda^{-1} f(\lambda x) \geq \lambda f(\lambda^{-1}\lambda x) = \lambda f(x). \tag{A.7}$$

Now let $a, b \in \mathbb{R}$. We can assume that $a \geq b > 0$. Then $\frac{b}{a} \in [0,1]$ and $(1 + \frac{b}{a}) \geq 1$.

$$\begin{aligned} f(a+b) &= f\Big(a\,\big(1+\tfrac{b}{a}\big)\Big) \\ &\geq \big(1+\tfrac{b}{a}\big) f(a) \qquad \text{(using (A.7))} \\ &= f(a) + \tfrac{b}{a}\, f(a) \\ &\geq f(a) + f(\tfrac{b}{a}\, a) \quad \text{(using (A.6))} \\ &= f(a) + f(b), \end{aligned}$$

which finishes the proof for convex f.
For concave f, the proof is analogous, just exchange the inequality signs in the above. □

Remark A.11.
Often the connection between convexity or concavity and sub- or superadditivity is the other way round. In the special case of Proposition A.10 convexity implies superadditivity. Instead, in more general settings a convex function will often be *sub*-additive. For example, any norm or seminorm on a linear space is convex but subadditive. Proposition A.10 holds only in 1D.

Corollary A.12 (x^p is superadditive).
Suppose $p \geq 1$. Then, for all $a, b \geq 0$

$$\begin{aligned} (a+b)^p &\geq a^p + b^p \qquad \textit{and} \\ (a+b)^{1/p} &\leq a^{1/p} + b^{1/p}. \end{aligned}$$

Furthermore, bounds in the other direction exist as well:

$$\begin{aligned} (a+b)^p &\leq 2^{p-1}(a^p + b^p) \qquad \textit{and} \\ (a+b)^{1/p} &\geq 2^{1/p-1}(a^{1/p} + b^{1/p}). \end{aligned}$$

All these inequalities are sharp.

Proof. By Corollary A.9 $x \mapsto x^p$ is convex and $x \mapsto x^{1/p}$ is concave on $\mathbb{R}_{\geq 0}$. Thus, the first two inequalities follow from Proposition A.10.
The last two inequalities are a direct consequence of the definition of convexity and concavity: For $x \mapsto x^p$, take $\theta = 1/2$ in (A.4) to get

$$\left(\frac{x+y}{2}\right)^p \leq \frac{x^p + y^p}{2}.$$

Similarly, for $x \mapsto x^{1/p}$, take $\theta = 1/2$ in (A.5) to get

$$\left(\frac{x+y}{2}\right)^{1/p} \geq \frac{x^{1/p} + y^{1/p}}{2}.$$

Rearranging these yields the two last inequalities. To see the sharpness choose $a = 1, b = 0$ in the first two inequalities and $a = b = 1$ in the last two. □

An easy induction proof yields the following result for convex functions, which is a generalisation of their defining property (A.4). It can also be viewed as a discrete version of Jensen's inequality (Lemma A.1).

Lemma A.13 (Generalising the property of convex functions).
Let X be a vector space, $K \subseteq X$ a convex subset and $f : K \to \mathbb{R}$ a convex function. Suppose that $n \geq 1$. Then for any $x_1, \ldots, x_n \in K$,

$$f\Big(\sum_{i=1}^{n} \theta_i x_i\Big) \leq \sum_{i=1}^{n} \theta_i f(x_i),$$

whenever $\theta_i \in [0,1]$ with $\sum_{i=1}^{n} \theta_i = 1$.

Proof. By induction. For $n = 1$, this is obvious. Assume that the lemma holds for a certain value of n and use this to prove it still holds for $n+1$: Let $\theta_i \in [0,1], i \in \{1, \ldots, n+1\}$ with $\sum_{i=1}^{n+1} \theta_i = 1$. Then we can write

$$f\Big(\sum_{i=1}^{n+1} \theta_i x_i\Big) = f\Big((1-\theta_{n+1}) \sum_{i=1}^{n} \frac{\theta_i}{1-\theta_{n+1}} x_i + \theta_{n+1} x_{n+1}\Big),$$

and, using the convexity (A.4) we get,

$$f\Big(\sum_{i=1}^{n+1} \theta_i x_i\Big) \leq (1-\theta_{n+1}) f\Big(\sum_{i=1}^{n} \frac{\theta_i}{1-\theta_{n+1}} x_i\Big) + \theta_{n+1} f(x_{n+1}).$$

Now we use the induction hypothesis to get, since $\sum_{i=1}^{n} \frac{\theta_i}{1-\theta_{n+1}} = 1$,

$$f\Big(\sum_{i=1}^{n+1} \theta_i x_i\Big) \leq (1-\theta_{n+1}) \sum_{i=1}^{n} \frac{\theta_i}{1-\theta_{n+1}} f(x_i) + \theta_{n+1} f(x_{n+1}) = \sum_{i=1}^{n+1} \theta_i f(x_i),$$

which shows that the lemma also holds for $n+1$. Thus by induction n can be any positive integer. □

Although Corollary A.12 is useful on its own, it is really a special case of the equivalence of finite-dimensional norms. Here is a generalisation of the first two inequalities of that corollary:

Proposition A.14 (Equivalence of p-norms in $\mathbb{R}^n$).
Suppose $1 \leq p \leq q \leq \infty$. Then, for all $\boldsymbol{x} \in \mathbb{R}^n$,

$$\|\boldsymbol{x}\|_1 \geq \|\boldsymbol{x}\|_p \geq \|\boldsymbol{x}\|_q \geq \|\boldsymbol{x}\|_\infty. \tag{A.8}$$

and for bounding in the other direction,

$$\|\boldsymbol{x}\|_p \leq n^{\frac{1}{p}-\frac{1}{q}} \|\boldsymbol{x}\|_q, \tag{A.9}$$

where we set $1/\infty = 0$. All these bounds are sharp.

Proof. First observe that Corollary A.12 extends inductively to sums of more than two summands. Let $1 \leq p \leq q < \infty$. Then $\frac{q}{p} \geq 1$.

$$\|\boldsymbol{x}\|_q = \left(\sum_{i=1}^{n} |x_i|^q\right)^{\frac{1}{q}} = \left(\sum_{i=1}^{n} |x_i|^{p\frac{q}{p}}\right)^{\frac{1}{q}} \leq \left(\left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{q}{p}}\right)^{\frac{1}{q}} = \|\boldsymbol{x}\|_p.$$

For proving (A.8), it remains to show that $\|\boldsymbol{x}\|_q \geq \|\boldsymbol{x}\|_\infty$ for $1 \leq q < \infty$:

$$\|\boldsymbol{x}\|_q = \left(\sum_{i=1}^{n} |x_i|^q\right)^{1/q} \geq \left(\max_{1\leq i\leq n} |x_i|^q\right)^{1/q} = \max_{1\leq i\leq n} (|x_i|^q)^{1/q} = \|\boldsymbol{x}\|_\infty.$$

To prove (A.9), we start with the case $p < q < \infty$. In this case $x \to x^{q/p}$ is convex, thus

$$\Big(\sum_{i=1}^{n} \frac{1}{n}|x_i|^p\Big)^{q/p} \leq \sum_{i=1}^{n} \frac{1}{n}(|x_i|^p)^{q/p}.$$

Taking the q-th root gives

$$\big(1/n\big)^{1/p}\|\boldsymbol{x}\|_p \leq \big(1/n\big)^{1/q}\|\boldsymbol{x}\|_q,$$

which yields (A.9) for this case. The remaining case is wlog $p < q = \infty$:

$$\|\boldsymbol{x}\|_p = \Big(\sum_{i=1}^{n}|x_i|^p\Big)^{1/p} \leq \Big(\sum_{i=1}^{n}\|\boldsymbol{x}\|_\infty^p\Big)^{1/p} = n^{1/p}\|\boldsymbol{x}\|_\infty.$$

To conclude the proof we note that the sharpness of the bounds follows by choosing $\boldsymbol{x} = (1, 1, \ldots, 1)$ in (A.9) and $\boldsymbol{x} = (1, 0, \ldots, 0)$ in (A.8). □

For convenience we state a special case of this proposition: For $n \geq 1$ and for all $\boldsymbol{x} \in \mathbb{R}^n$, we have

$$\big(\sum_{i=1}^{n} x_i\big)^2 \leq n\sum_{i=1}^{n} x_i^2, \tag{A.10}$$

which follows by setting $p = 1$ and $q = 2$ in (A.9) and squaring both sides. Observe that this inequality also holds for negative x_i.

A.4 Gelfand triple

This section is an introduction to Gelfand triples, giving the necessary background to define them. The theory is taken from the following books: Wloka [Wlo87, §17], Oden/Demkowicz [OD96] and Renardy/Rogers [RR04].

Let X, Y be normed vector spaces. Denote by $\mathscr{L}(X, Y)$ the vector space of continuous linear mappings from X to Y.

Definition A.15 (Dual space).
Let X be a Banach space. Define the *dual space* of X by

$$X' := \mathscr{L}(X, \mathbb{R}).$$

Associated with X' is the *duality pairing* $\langle\cdot,\cdot\rangle_X : X' \times X \to \mathbb{R}$ defined through

$$\langle y, x\rangle_X := y(x),$$

for any $y \in X'$ and $x \in X$, and the norm

$$\|y\|_{X'} := \sup_{0\neq x\in X} \frac{|\langle y, x\rangle_X|}{\|x\|_X}, \tag{A.11}$$

for any $y \in X'$.

Before we can continue, we state the theorem of Hahn-Banach:

Theorem A.16 (Hahn-Banach).
Let X be a Banach space and let M be a linear subspace of X. Let l_M be a bounded linear mapping from M to $\mathbb{R}$. Then there is a linear functional $l \in X'$ such that $l|_M = l_M$ and $\|l\|_{X'} = \|l_M\|_{M'} := \sup_{0 \neq x \in M} \frac{|l_M(x)|}{\|x\|_X}$.

Proof. This formulation is taken from [RR04, Theorem 6.56]. A more general version, for *sublinear* functionals is proved in [OD96, Section 5.4]. □

In the definition of the norm of the dual space (A.11), the rôle of X and X' can actually be exchanged, as the following lemma shows:

Lemma A.17 (An equality with the norm of the dual space).
Let X be a normed vector space. Then

$$\|x\|_X = \sup_{0 \neq x' \in X'} \frac{|\langle x', x \rangle_X|}{\|x'\|_{X'}}.$$

Proof. Firstly, immediately from (A.11),

$$\|x\|_X \geq \frac{|\langle x', x \rangle_X|}{\|x'\|_{X'}}. \tag{A.12}$$

Now let $x \in X$ be fixed. Set $M = \{\lambda x : \lambda \in \mathbb{R}\}$ and define $l_M : M \to \mathbb{R}$ by $l_M(\lambda x) := \lambda \|x\|_X$ for $\lambda \in \mathbb{R}$. We see that l_M is linear and bounded. By the theorem of Hahn-Banach, Theorem A.16, l_M can be extended to some $x' \in X'$ such that $\|x'\|_{X'} = 1$ and $\langle x', x \rangle_X = \|x\|_X$. Thus for each $x \in X$ there is an $x' \in X'$ such that we get equality in (A.12), which completes the proof. □

Definition A.18 (Continuous and dense embedding).
Let X, Y be normed vector spaces. Then we write

$$X \hookrightarrow Y,$$

if there is a continuous linear embedding from X into Y i.e. if there is an injective $\iota \in \mathscr{L}(X, Y)$. Furthermore, if $\iota(X)$ is dense in Y, then the embedding is also called *dense*.

Remark A.19 (Continuous embeddings of function spaces).
Some authors define a continuous embedding only for the case that $X \subseteq Y$, for example [RR04, Definition 7.15]. When the definition of continuous embeddings is applied to spaces of functions, one further requirement is always implicitly there: Any function $v \in X$ and its embedded image $\iota(v)$ should be the *same* function (but again identifying functions which differ on sets of measure 0), i.e. ι should be the identity mapping. Without this further assumption, saying that $H^1(\Omega)$ is continuously embedded in $L_2(\Omega)$ would not mean much, since both are separable infinite-dimensional Hilbert spaces and thus they are isometrically isomorphic, an even stronger statement.

Definition A.20 (Dual mapping).
Let X, Y be normed vector spaces and let $A \in \mathscr{L}(X, Y)$. The *dual* of A, denoted by A' is the mapping

$$A' : \left\{ \begin{array}{ccc} Y' & \to & X' \\ y' & \mapsto & A(y'), \end{array} \right.$$

where $A(y')$ is defined via

$$\langle\, A'(y'), x \,\rangle_X = \langle\, y', A(x) \,\rangle_Y.$$

for any $x \in X$.

Lemma A.21 (Properties of dual mapping).
Let $A \in \mathscr{L}(X,Y)$, then $A' \in \mathscr{L}(Y',X')$. Moreover $\|A'\|_{\mathscr{L}(Y',X')} = \|A\|_{\mathscr{L}(X,Y)}$.

Proof. A' is linear by the properties of vector spaces. The continuity of A' follows from the continuity of A:

$$\begin{aligned}\infty > \|A\|_{\mathscr{L}(X,Y)} &= \sup_{0\neq x\in X} \frac{\|A(x)\|_Y}{\|x\|_X} \overset{\text{Lemma A.17}}{=} \sup_{0\neq y'\in Y'} \sup_{0\neq x\in X} \frac{|\langle y', A(x)\rangle_Y|}{\|y'\|_{Y'}\|x\|_X} \\ &= \sup_{0\neq y'\in Y'} \sup_{0\neq x\in X} \frac{|\langle A'(y'), x\rangle_X|}{\|y'\|_{Y'}\|x\|_X} = \sup_{0\neq y'\in Y'} \frac{\|A'(y')\|_{X'}}{\|y'\|_{Y'}} = \|A'\|_{\mathscr{L}(Y',X')}.\end{aligned}$$

This also shows the equality of the norms of A and A'. □

Lemma A.22 (Injectivity of dual mapping).
Let X, Y be Banach spaces and $A \in \mathscr{L}(X,Y)$. Then

$$A(X) \text{ dense in } Y \iff A' \text{ is injective.}$$

Proof. This can be proved using the Hahn-Banach theorem (Theorem A.16), see [Wlo87, Theorem 17.1] or [DL88, VI.1.4 Corollary 9] □

A Banach space X is called *reflexive*, if the dual of its dual space is again the space X:

$$(X')' = X.$$

Lemma A.23 (Embedding of dual spaces).
Let X, Y be reflexive Banach spaces such that $X \hookrightarrow Y$ with continuous, dense embedding. Then $Y' \hookrightarrow X'$ with continuous, dense embedding. Furthermore, if we identify the spaces with their embedded images, then for any $x \in X$ and $y' \in Y'$

$$\langle y', x\rangle_X = \langle y', x\rangle_Y.$$

Proof. Let $\iota : X \to Y$ be the continuous, dense embedding, and denote by $\langle\cdot,\cdot\rangle_X$ the duality pairing of X' and X, analogously for Y. The dual of ι, $\iota' : Y' \to X'$ is defined through

$$\langle \iota'(y'), x\rangle_X := \langle y', \iota(x)\rangle_Y,$$

for any $y' \in Y'$. Now, by Lemma A.22, the fact that $\iota(X)$ is dense in Y implies that ι' is injective. Thus it remains to show that $\iota'(Y)$ is dense in X'. The other direction of Lemma A.22 assures that $\iota'(Y)$ is dense in X' iff ι'' is injective. As the defining equation for ι'' is

$$\langle \iota''(x''), y'\rangle_{Y'} = \langle x'', \iota'(y')\rangle_{X'},$$

and since X and Y are reflexive, we have $\langle x', x\rangle_{X'} = \langle x', x\rangle_{X''}$, thus

$$\langle y', \iota''(x)\rangle_Y = \langle \iota'(y'), x\rangle_X,$$

whence $\iota'' = \iota$. The identity of the duality pairings is just the definition of ι' from above after omitting the embeddings. □

Theorem A.24 (Riesz' representation theorem).
Let $(H, (\cdot,\cdot)_H)$ be a Hilbert space. Then there is an isometric isomorphism $R : H \to H'$ such that

$$(h, v)_H = \langle R(h), v\rangle_H \qquad \forall v \in H.$$

Remark A.25.
Note that the mapping R from Theorem A.24 is linear since it is an isomorphism.

Proof of Theorem A.24. The statement and proof of this can be found for example in [OD96, Theorem 6.4.1, Corollary 6.4.1] and in [RR04, Theorem 6.5.2]. □

A consequence of the Riesz representation theorem is that all Hilbert spaces are reflexive.

Proposition A.26 (Gelfand triple).
Let $(H, (\cdot,\cdot)_H)$ and $(V, (\cdot,\cdot)_V)$ be Hilbert spaces such that $V \hookrightarrow H$ with continuous dense embedding. Then, by identifying H with its dual H',

$$V \hookrightarrow H \hookrightarrow V'$$

with continuous, dense embeddings. Furthermore, let $\langle\cdot,\cdot\rangle_V$ denote the duality pairing between V' and V. Then

$$\langle h, v\rangle_V = (h, v)_H \quad \forall\, h \in H, v \in V.$$

Proof. Apply Lemma A.23 to V and H and use the Riesz isometry $H \to H'$, see Theorem A.24. □

A Gelfand triple is also sometimes called a *Hilbert triplet*, see [BG87b, Section 1.4G].

Remark A.27.
By applying the Riesz representation theorem to V, the elements of V' could also be represented by elements of V. But this representation is not used, since in that case some elements of V' would be represented by two different functions, one from the isometry $V \to V'$ and another one from the embedding $H \to V'$. For more background, see [BG87b, Section 1.4G] and [Wlo87, Section 17.1].

A.5 Positive definite matrices

This section is intended as a supplement to Subsection 5.4.2, which looks at the question of whether the discrete solution for $Q = 3$ is uniquely determined. Here we give some counter-examples and background on positive definite matrices.

First we recall the definition of positive definite matrices.

Definition A.28 (Positive definite matrix).
Let $\underline{\boldsymbol{A}}$ be a square $n \times n$-matrix. Then $\underline{\boldsymbol{A}}$ is called *positive definite* if and only if for all $\boldsymbol{x} \in \mathbb{R}^n \setminus \{\boldsymbol{0}\}$

$$\boldsymbol{x}^T \underline{\boldsymbol{A}} \boldsymbol{x} > 0.$$

Note that this definition also applies to non-symmetric matrices.

For symmetric matrices an important criterion is the following: A symmetric matrix is positive definite if and only if all its eigenvalues are positive. For non-symmetric matrices however this criterion cannot be applied directly, as the next example shows.

Example A.29 (Counterexample with a non-symmetric matrix).
We give an example of a matrix which has only positive eigenvalues but which is not positive definite. Let

$$\underline{\boldsymbol{A}} := \begin{pmatrix} 2 & 3 \\ 0 & 1 \end{pmatrix},$$

which has the two positive eigenvalues 1 and 2. On the other hand, choosing $\boldsymbol{x}^T = (2, -3)$, we obtain

$$\boldsymbol{x}^T \underline{\boldsymbol{A}} \boldsymbol{x} = (2, -3) \begin{pmatrix} 2 & 3 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 2 \\ -3 \end{pmatrix} = -1.$$

Thus $\underline{\boldsymbol{A}}$ is not positive definite.

In Remark 5.10, we would have liked to show that the product $\underline{\boldsymbol{A}}\underline{\boldsymbol{B}}^n$ of the stiffness matrix with a diagonal matrix is positive definite. The next lemma gives a counter-example to a conjecture which would have helped.

Lemma A.30 (A counter-example).
Let $\underline{\boldsymbol{A}}$ be a symmetric positive definite $n \times n$-matrix and $\underline{\boldsymbol{B}}$ a positive definite diagonal $n \times n$-matrix. Then their product $\underline{\boldsymbol{A}}\underline{\boldsymbol{B}}$ is not necessarily *positive definite.*

Proof. It is sufficient to give a counter-example. Let

$$\underline{\boldsymbol{A}} := \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \quad \text{and} \quad \underline{\boldsymbol{B}} := \begin{pmatrix} 15 & 0 \\ 0 & 1 \end{pmatrix}.$$

Now $\underline{\boldsymbol{A}}$ is positive definite since for any $\boldsymbol{x} \in \mathbb{R}^n \setminus \{\boldsymbol{0}\}$,

$$\boldsymbol{x}^T \underline{\boldsymbol{A}} \boldsymbol{x} = 2x_1^2 + 2x_2^2 + 2x_1x_2 = x_1^2 + x_2^2 + (x_1 + x_2)^2 > 0.$$

The product of the two matrices is

$$\underline{\boldsymbol{A}}\underline{\boldsymbol{B}} = \begin{pmatrix} 30 & 1 \\ 15 & 2 \end{pmatrix}.$$

With the choice of $\boldsymbol{x}^T = (-1, 4)$, we obtain

$$\boldsymbol{x}^T \underline{\boldsymbol{A}}\underline{\boldsymbol{B}} \boldsymbol{x} = (-1, 4) \begin{pmatrix} 30 & 1 \\ 15 & 2 \end{pmatrix} \begin{pmatrix} -1 \\ 4 \end{pmatrix} = -2.$$

Thus $\underline{\boldsymbol{A}}\underline{\boldsymbol{B}}$ is not positive definite. □

Here is what can be proved in the situation of the previous lemma, using linear algebra.

Lemma A.31 (Product of symmetric positive definite matrices).
Let $\underline{\boldsymbol{A}}$ be a symmetric positive definite $n \times n$-matrix and $\underline{\boldsymbol{B}}$ a positive definite diagonal $n \times n$-matrix. Then their product $\underline{\boldsymbol{A}}\underline{\boldsymbol{B}}$ has only positive eigenvalues.

Proof. Since $\underline{\boldsymbol{B}}$ has only positive entries in the diagonal, the matrix $\underline{\boldsymbol{B}}^{1/2}$ exists and is invertible. Moreover $\underline{\boldsymbol{A}}$ is symmetric. Thus, by Sylvester's law of inertia, the matrix $(\underline{\boldsymbol{B}}^{1/2})^T \underline{\boldsymbol{A}}\underline{\boldsymbol{B}}^{1/2}$ has the same number of positive eigenvalues as $\underline{\boldsymbol{A}}$ has, i.e. all n of them. Therefore $(\underline{\boldsymbol{B}}^{1/2})^T \underline{\boldsymbol{A}}\underline{\boldsymbol{B}}^{1/2}$ is positive definite. Now we can write

$$(\underline{\boldsymbol{B}}^{1/2})^T \underline{\boldsymbol{A}}\underline{\boldsymbol{B}}^{1/2} = \underline{\boldsymbol{B}}^{1/2}(\underline{\boldsymbol{A}}\underline{\boldsymbol{B}})\underline{\boldsymbol{B}}^{-1/2}.$$

This implies that $(\underline{\boldsymbol{B}}^{1/2})^T \underline{\boldsymbol{A}}\underline{\boldsymbol{B}}^{1/2}$ is similar to $\underline{\boldsymbol{A}}\underline{\boldsymbol{B}}$, whence both matrices have the same spectrum of eigenvalues. So the eigenvalues of $\underline{\boldsymbol{A}}\underline{\boldsymbol{B}}$ are all positive. □

By comparing with Example A.29, we see that $\underline{\boldsymbol{A}}\underline{\boldsymbol{B}}$ need not be positive definite, since it is in general non-symmetric.

Appendix B

Amann's Result of Existence and Uniqueness

In this chapter of the appendix we would like to demonstrate how Amann's result in [Ama91] applies to the Cohen model in 2D and 3D. Amongst the articles on existence and uniqueness cited in Section 3.3, this is the most relevant to our form of the equations. The article is motivated by the one-dimensional form of the model but for his main theorem Amann generalises not only the space dimension, but also the differential operators and boundary conditions. Therefore we attempt to specialise the theorem again to the Cohen model and arrive at the following corollary.

For this appendix, suppose that Ω is a bounded domain in $\mathbb{R}^d$ such that $\bar{\Omega}$ is a d-dimensional C^∞-submanifold of $\mathbb{R}^d$. We also set $\mathbb{R}^+ := \{x \in \mathbb{R} \,:\, x \geq 0\}$.

Corollary B.1 (Amann's result specialised for the Cohen model). *Fix $p \in [2, \infty)$ with $p > d$. Suppose that*

$$\begin{aligned} f, \hat{f} &\in C^\infty(\overline{\Omega} \times \mathbb{R}^+), \\ u_D &\in C^\infty(\overline{\Omega} \times \mathbb{R}^+), \\ \breve{u} &\in W_p^2(\Omega) \text{ with } \breve{u} = u_D(\cdot, 0) \text{ on } \partial\Omega, \\ \breve{\sigma} &\in W_p^2(\Omega). \end{aligned}$$

Then the problem of finding u and σ with

$$\begin{aligned} \dot{u} - \Delta u - \Delta\sigma &= f && \text{in } \Omega \times (0, \infty), \\ \dot{\sigma} + \gamma(u)\sigma &= u + \hat{f} && \text{in } \Omega \times (0, \infty), \\ u &= u_D && \text{on } \partial\Omega \times (0, \infty), \\ u(0) &= \breve{u} && \text{on } \Omega, \\ \sigma(0) &= \breve{\sigma} && \text{on } \Omega, \end{aligned} \tag{B.1}$$

has a unique solution

$$(u, \sigma) \in C(\mathbb{R}^+, W_p^2(\Omega, \mathbb{R}^2)) \cap C^1(\mathbb{R}^+, L_p(\Omega, \mathbb{R}^2)).$$

In order to verify that it is indeed a special case of Amann's theorem from [Ama91], we cite it here for convenience, adapting notation slightly. We start by repeating the definitions from that article:

Let $\mathcal{A}_1$ and $\mathcal{A}_2$ be two differential operators on Ω, with real coefficients,

$$\mathcal{A}_r u := -\sum_{j,k=1}^{d} \partial_j(a_{jk}^r \partial_k u) + \sum_{j=1}^{d} a_j^r \partial_j u + a_0^r u, \quad r = 1, 2,$$

such that

$$a^r_{jk} = a^r_{kj}, a^r_j, a^r_0 \in C^\infty(\bar\Omega), j,k = 1,\dots,d, r = 1,2.$$

and

$$\sum_{j,k=1}^{d} a^1_{jk}(x)\xi^j\xi^k > 0, x \in \bar\Omega, \xi = (\xi^1,\dots,\xi^d) \in \mathbb{R}^d \setminus \{0\}.$$

Let also $\delta \in C(\partial\Omega, \{0,1\})$ be fixed and define a boundary operator $\mathcal{B}$ on $\partial\Omega$ by

$$\mathcal{B}u := \delta\Big(\sum_{j,k=1}^{d} a^1_{jk}\boldsymbol{n}_j\partial_k u + bu\Big) + (1-\delta)u,$$

where $b \in C^\infty(\partial\Omega)$ is real-valued and $\boldsymbol{n} = (\boldsymbol{n}_1,\dots,\boldsymbol{n}_d)$ is the unit outer normal on $\partial\Omega$. Observe that δ is constant on connected components of $\partial\Omega$ and that $\mathcal{B}$ is simply the Dirichlet boundary operator on those components of $\partial\Omega$ on which $\delta = 0$, whereas it is a Neumann type boundary operator on the remaining components.
Given $q \in (1,\infty)$, we put

$$H^2_{q,\mathcal{B}}(\Omega) := \{v \in H^2_q(\Omega) \,:\, \mathcal{B}v = 0\}, \tag{B.2}$$

where $H^2_q(\Omega)$ is the Sobolev space $W^2_q(\Omega)$ since $q < \infty$. Furthermore,

$$A_1 := \mathcal{A}_1|H_{q,\mathcal{B}}, \tag{B.3}$$

so that A_1 is the "L_q-realisation of the elliptic boundary value problem $(\mathcal{A}_1,\mathcal{B})$". It is assumed without loss of generality that 0 is in the resolvent set of A_1. Finally, $\mathcal{A}^\#_2$ is the formal adjoint of $\mathcal{A}_2$, that is,

$$\mathcal{A}^\#_2 v := -\sum_{j,k=1}^{d} \partial_j(a^2_{jk}\partial_k v) - \sum_{j=0}^{d} \partial_j(a^2_j v) + a^2_0 v, \tag{B.4}$$

and A_1' is the dual of A_1.

Theorem B.2 (Amann).
Suppose that

$$f := (f_1, f_2) \in C^\infty(\bar\Omega \times \mathbb{R}^+ \times \mathbb{R}^2, \mathbb{R}^2).$$

Given any $p \in (d,\infty)$ with $p \geq 2$, then the problem of finding $u := (u_1,u_2)$ such that

$$\begin{aligned} \dot u_1 + \mathcal{A}_1u_1 + \mathcal{A}_2u_2 &= f_1(x,t,u) && \text{in } \Omega\times(0,\infty),\\ \dot u_2 &= f_2(x,t,u) && \text{in } \Omega\times(0,\infty),\\ \mathcal{B}u_1 &= 0 && \text{on } \partial\Omega\times(0,\infty),\\ u(\cdot,0) &= u_0 && \text{on } \Omega, \end{aligned} \tag{B.5}$$

possesses for each $u_0 \in H^2_{p,\mathcal{B}}(\Omega)\times H^2_p(\Omega)$ a unique solution

$$u(.,u_0) \in C(\mathbb{R}^+, H^2_p(\Omega,\mathbb{R}^2)) \cap C^1(\mathbb{R}^+, L_p(\Omega,\mathbb{R}^2)), \tag{B.6}$$

provided that there exists a constant c such that

$$\|\mathcal{A}^\#_2(A_1')^{-1}\|_{\mathscr{L}(L_q(\Omega),L_q(\Omega))} \leq c, \quad 1 < q \leq 2, \tag{B.7}$$

and that there is a function $\alpha \in C(\mathbb{R}^+,\mathbb{R}^+)$ such that

$$|f(x,t,\eta)| \leq \alpha(t)(1+|\eta|), \quad (x,t,\eta) \in \bar\Omega\times\mathbb{R}^+\times\mathbb{R}^2. \tag{B.8}$$

Proof of Corollary B.1. We can directly apply Theorem B.2 to $u_1 := u - u_D$, $u_2 := \sigma$. Choose the Dirichlet boundary conditions $\delta \equiv 0$, i.e.

$$\mathcal{B}u_1 := u_1 = 0.$$

Also, choose the differential operators

$$\mathcal{A}_1 := \mathcal{A}_2 := -\Delta,$$

whence the formal adjoint of $\mathcal{A}_2$ is $\mathcal{A}_2^{\#} = -\Delta$ and the "L_p-realization" of $\mathcal{A}_1$ is $A_1 : H^2_{p,\mathcal{B}} \to L_p$. The dual of A_1 is $A_1' = -\Delta : H^2_{q,\mathcal{B}} \to L_q(\Omega)$ with $1/p + 1/q = 1$, since $(v, -\Delta w) = (-\Delta v, w)$ for all $v \in H^2_{q,\mathcal{B}}$, $w \in H^2_{p,\mathcal{B}}$. Thus $\mathcal{A}_2^{\#}(A_1')^{-1}$ is the identity on $L_q(\Omega)$. Hence

$$\|\mathcal{A}_2^{\#}(A_1')^{-1}\|_{\mathscr{L}(L_q(\Omega),L_q(\Omega))} = 1,$$

and condition (B.7) of the theorem is shown.

Furthermore, choose

$$\begin{aligned} f_1(\boldsymbol{x}, t, (\eta_1, \eta_2)) &:= f(\boldsymbol{x}, t) - \dot{u}_D(\boldsymbol{x}, t) + \Delta u_D(\boldsymbol{x}, t) \\ f_2(\boldsymbol{x}, t, (\eta_1, \eta_2)) &:= u_D(\boldsymbol{x}, t) + \eta_1 - \gamma(u_D(\boldsymbol{x}, t) + \eta_1)\eta_2 + \hat{f}(\boldsymbol{x}, t), \end{aligned}$$

and of course the initial condition $u_0 := (\breve{u} - u_D(\cdot, 0), \breve{\sigma})$. To show condition (B.8) of the theorem:

$$\begin{aligned} &|(f_1(\boldsymbol{x}, t, (\eta_1, \eta_2), f_2(\boldsymbol{x}, t, (\eta_1, \eta_2))| \\ &\quad \le |f(\boldsymbol{x}, t)| + |\dot{u}_D(\boldsymbol{x}, t)| + |\Delta u_D(\boldsymbol{x}, t)| + |u_D(\boldsymbol{x}, t)| + |\eta_1| + \hat{\gamma}|\eta_2| + |\hat{f}(\boldsymbol{x}, t)| \\ &\qquad\qquad \le \alpha(t)(1 + |(\eta_1, \eta_2)|), \end{aligned}$$

where

$$\begin{aligned} \alpha(t) := 2\max(1, \hat{\gamma})\Big(&\|f(\cdot, t)\|_\infty + \|\hat{f}(\cdot, t)\|_\infty \\ &+ \|u_D(\cdot, t)\|_\infty + \|\dot{u}_D(\cdot, t)\|_\infty + \|\Delta u_D(\cdot, t)\|_\infty \Big). \end{aligned}$$

□

References

[Ama90] Herbert Amann. *Ordinary Differential Equations: An Introduction to Nonlinear Analysis*. Number 13 in de Gruyter Studies in Mathematics. Walter de Gruyter, Berlin, 1990.

[Ama91] Herbert Amann. Global existence for a class of highly degenerate parabolic systems. *Japan J. Indust. Appl. Math.*, 8(1):143–151, 1991.

[Ama93] Herbert Amann. Nonhomogeneous linear and quasilinear elliptic and parabolic boundary value problems. In H.J. Schmeisser and H. Triebel, editors, *Function Spaces, Differential Operators and Nonlinear Analysis*, pages 9–126. Teubner, Stuttgart, Leipzig, 1993. `www.math.unizh.ch/amann`.

[BG87a] Franco Brezzi and Gianni Gilardi. *Finite Element Handbook*, chapter 2: Functional Spaces. In Kardestuncer [Kar87], 1987.

[BG87b] Franco Brezzi and Gianni Gilardi. *Finite Element Handbook*, chapter 1: Functional Analysis. In Kardestuncer [Kar87], 1987.

[BGN03] John W. Barrett, Harald Garcke, and Robert Nürnberg. Finite element approximation of surfactant spreading on a thin film. *SIAM J. Numer. Anal.*, 41(4):1427–1464, 2003.

[Bou04] Nicolas Bourbaki. *Elements of Mathematics: Functions of a Real Variable*. Springer-Verlag, Berlin, 2004.

[Bra97] Dietrich Braess. *Finite Elements: Theory, Fast Solvers and Applications in Solid Mechanics*. Cambridge University Press, Cambridge, 1997.

[BS94] Susanne C. Brenner and L. Ridgway Scott. *The Mathematical Theory of Finite Element Methods*. Number 15 in Texts in Applied Mathematics. Springer-Verlag, New York, 1994.

[CGMSS81] Ian Christie, David F. Griffiths, A. Ronald Mitchell, and Jesús María Sanz-Serna. Product approximation for nonlinear problems in the finite element method. *IMA J. Numer. Anal.*, 1(3):253–266, 1981.

[Cia02] Philippe G. Ciarlet. *The Finite Element Method for Elliptic Problems*. Number 40 in Classics in Applied Mathematics. SIAM, Philadelphia, 2002.

[CW91] Donald S. Cohen and Andrew B. White, Jr. Sharp fronts due to diffusion and viscoelastic relaxation in polymers. *SIAM J. Appl. Math.*, 51(2):472–483, 1991.

[CWW95] Donald S. Cohen, Andrew B. White, Jr., and Thomas P. Witelski. Shock formation in a multidimensional viscoelastic diffusive system. *SIAM J. Appl. Math.*, 55(2):348–368, 1995.

[DL88] Robert Dautray and Jacques-Louis Lions. *Mathematical Analysis and Numerical Methods for Science and Technology*. Springer-Verlag, Berlin, first edition, 1988. 6 volumes.

[Edw95] David A. Edwards. Constant front speed in weakly diffusive non-fickian systems. *SIAM Journal on Applied Mathematics*, 55(4):1039–1058, 1995.

[EG04] Alexandre Ern and Jean-Luc Guermond. *Theory and Practice of Finite Elements*. Number 159 in Applied Mathematical Sciences. Springer-Verlag, New York, 2004.

[Eva98] Lawrence C. Evans. *Partial Differential Equations*. Number 19 in Graduate Studies in Mathematics. AMS, Providence, Rhode Island, 1998.

[GLS90] Gustaf Gripenberg, Stig-Olof Londen, and Olof Staffans. *Volterra Integral and Functional Equations*. Number 34 in Encyclopedia of Mathematics and its Applications. Cambridge University Press, Cambridge, 1990.

[Hac95] Wolfgang Hackbusch. *Integral Equations: Theory and Numerical Treatment*. Number 120 in International Series of Numerical Mathematics. Birkhäuser, Basel, 1995.

[Heu00] Harro Heuser. *Lehrbuch der Analysis; Teil 2*. B.G. Teubner, Stuttgart, eleventh edition, 2000.

[Heu01] Harro Heuser. *Lehrbuch der Analysis; Teil 1*. B.G. Teubner, Stuttgart, fourteenth edition, 2001.

[HZ96] Bei Hu and Jianhua Zhang. Global existence for a class of non-Fickian polymer-penetrant systems. *J. Partial Differential Equations*, 9(3):193–208, 1996.

[Kar87] Hayrettin Kardestuncer, editor. *Finite Element Handbook*. McGraw-Hill, New York, 1987.

[NLM$^+$] David J. Needham, John A. Leach, John Merkin, Norbert Bauermeister, and Simon Shaw. On the properties of degenerate nonlinear non-Fickian polymer diffusion systems. In preparation.

[OD96] J. Tinsley Oden and Leszek F. Demkowicz. *Applied Functional Analysis*. CRC Press, Boca Raton, 1996.

[PTVF02] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in C*. Cambridge University Press, second edition, 2002.

[RR04] Michael Renardy and Robert C. Rogers. *An Introduction to Partial Differential Equations*. Number 13 in Texts in Applied Mathematics. Springer-Verlag, New York, second edition, 2004.

[RS05] Béatrice Rivière and Simon Shaw. Discontinuous Galerkin finite element approximation of nonlinear non-Fickian diffusion in viscoelastic polymers. Technical Report BICOM 05/6, BICOM, Brunel University, London, 2005.

[RS06] Béatrice Rivière and Simon Shaw. Discontinuous Galerkin finite element approximation of nonlinear non-Fickian diffusion in viscoelastic polymers. *SIAM J. Numer. Anal.*, 44(6):2650–2670, 2006.

[SE04] Sumanth Swaminathan and David A. Edwards. Traveling waves for anomalous diffusion in polymers. *Applied Mathematics Letters*, 17(1):7–12, 2004.

[SFB97] SFB256, Bonn. *GRAPE Manual Version 5.3*, 1997. `www.iam.uni-bonn.de/sfb256/grape/`.

[SS05] Alfred Schmidt and Kunibert G. Siebert. *Design of Adaptive Finite Element Software*. Number 42 in Lecture Notes in Computational Science and Engineering. Springer-Verlag, Berlin, 2005. (see `www.alberta-fem.de`).

[SWW04] Simon Shaw, M. K. Warby, and J. R. Whiteman. An introduction to the theory and numerical analysis of viscoelasticity problems, September 2004. (private communication).

[Tem97] Roger Temam. *Infinite-Dimensional Dynamical Systems in Mechanics and Physics*. Number 68 in Applied Mathematical Sciences. Springer-Verlag, New York, second edition, 1997.

[TW78] Noreen Thomas and A. H. Windle. Transport of methanol in poly(methyl methacrylate). *Polymer*, 19(3):255–265, 1978.

[Vor08] Dmitry A. Vorotnikov. Dissipative solutions for equations of viscoelastic diffusion in polymers. *J. Math. Anal. Appl.*, 339(2):876–888, 2008.

[Wlo87] Joseph Wloka. *Partial Differential Equations*. Cambridge University Press, Cambridge, 1987.

[Zhe04] Songmu Zheng. *Nonlinear Evolution Equations*. Number 133 in Monographs and Surveys in Pure and Applied Mathematics. Chapman & Hall/CRC, Boca Raton, FL, 2004.

List of Symbols

Basic Notation

$:=$	equation defines the left-hand side
$\hookrightarrow$	continuous embedding, page 181
$\boldsymbol{x}^T\boldsymbol{y}$	Euclidean inner product of $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, page 16
$\underline{\boldsymbol{A}}^T$	transpose of matrix $\underline{\boldsymbol{A}}$
$\mathbb{N}$	set of natural numbers $\{0, 1, 2, \ldots\}$, page 16
$\mathbb{R}$	set of real numbers, page 16
$\mathbb{R}^+$	$:= \{x \in \mathbb{R} \,:\, x \geq 0\}$, page 185
$\varnothing$	the empty set $\{\}$
$A \setminus B$	set difference, also defined for $B \not\subset A$: $A \setminus B := \{x \in A : x \notin B\}$
(a, b)	open interval $\{x \in \mathbb{R} : a < x < b\}$
$[a, b]$	closed interval $\{x \in \mathbb{R} : a \leq x \leq b\}$
δ_{ij}	Kronecker delta: $\delta_{ij} = 1$ if $i = j$ and 0 otherwise
$\vee$	logical OR
$\wedge$	logical AND
$\exists$	there exists
$\forall$	for all
$\lvert A \rvert$	cardinality of set A
∇	Nabla operator
Δ	Laplace operator, ∇^2
$\mathcal{G}$	inverse Laplacian, see equation (2.16), page 22
$\dot{v}$	first time derivative, page 22
$\ddot{v}$	second time derivative, page 22
$\dddot{v}$	third time derivative, page 22
ess sup	essential supremum: $\operatorname{ess\,sup}_{x \in X} a(x) = \inf\{s \in \mathbb{R} : a(x) \leq s \text{ for a.e. } x \in X\}$
a.e.	almost every(where), that is for all values, except for some which form a set of measure 0

Constants

C	generic constant, that may have different values in different places, page 16
c_i	constants defined with explicit expressions in Chapters 4 and 6, page 16
C''_γ	bound for γ'', see equation (3.4), page 34
C'_γ	upper bound for γ', see equation (3.3), page 34

C_V	Constant such that $\|v\|_0 \leq C_V \|v\|_V$, see equation (2.11), page 20
$C_{0,\infty}$	Constant such that $\|v\|_0 \leq C_{0,\infty} \|v\|_{L_\infty(\Omega)}$ for $v \in L_\infty(\Omega)$, page 56
$C_{\infty,V}$	Constant such that $\|v\|_{L_\infty(\Omega)} \leq C_{\infty,V} \|v\|_V$ for $v \in V$, if $d = 1$, page 55

Function Spaces, Norms, Inner Products

$\mathscr{L}(X, Y)$	vector space of continuous linear mappings from X to Y, page 180
V	test space with functions which are vanishing on Γ_D, see equation (2.9), page 20
X'	dual space of Banach space X, page 180
$C^0(\bar{\Omega})$	space of continuous functions $\bar{\Omega} \to \mathbb{R}$, page 18
$L_p(\Omega)$	Lebesgue space, page 18
$W_p^m(\Omega)$	Sobolev space, page 18
$L_2(\Omega)^2$	$L_2(\Omega) \times L_2(\Omega)$, see equation (2.7), page 19
$\langle \cdot, \cdot \rangle$	duality pairing of V. $\langle \cdot, \cdot \rangle : V' \times V \to \mathbb{R}$, page 21
$\|\cdot\|_0$	L_2-norm, see equation (2.6), page 19
$\|\cdot\|_V$	norm of V, see equation (2.10), page 20
$(\cdot, \cdot)_V$	inner product of V, see equation (2.10), page 20
$(\cdot, \cdot)_0$	L_2 inner product, see equation (2.6), page 19
$\|\cdot\|_{C([a,b];X),L}$	weighted norm on the space $C([a, b]; X)$, see equation (4.16), page 48
$C^m([a, b]; X)$	Space of m times continuously differentiable functions from $[a, b]$ into X, page 23
$C([a, b]; X)$	Space of continuous functions $[a, b] \to X$, page 23
$H^m(a, b; X)$	A space for time-dependent functions, page 24
$L_p(a, b; X)$	A space for time-dependent functions, page 24
$\mathcal{W}(0, T; V, V')$	$= \{u : u \in L_2(0, T; V), \dot{u} \in L_2(0, T; V')\}$, page 24
$W_p^m(a, b; X)$	A space for time-dependent functions, page 23

Problem Variables and Parameters

D	parameter of unscaled Cohen model, see equation (3.1), page 33
d	dimension of domain Ω, $d \in \{1, 2, 3\}$, page 17
Δ_a	size of transition region of γ, page 37
f	'forcing' term in u-equation, see equation (3.1), page 33
$\hat{f}$	'forcing' term in σ-equation, see equation (3.1), page 33
g	inhomogeneous Neumann boundary conditions, see equation (3.1), page 33
γ	nonlinear function, page 34
γ_1	standard choice for γ, involving the tanh function, see equation (3.5), page 37
γ_2	second choice for γ, involving a polynomial, see equation (3.6), page 38
γ_3	third choice for γ, involving the sine function, see equation (3.7), page 39
$\hat{\gamma}$	upper bound for γ, see equation (3.2), page 34
$\check{\gamma}$	lower bound for γ, see equation (3.2), page 34
Γ_D	Dirichlet part of $\partial\Omega$, see equation (2.1), page 17
Γ_N	Neumann part of $\partial\Omega$, see equation (2.1), page 17
K	parameter of unscaled Cohen model, see equation (3.1), page 33

L	linear form on the right-hand side of u-equation, see equation (4.2), page 42
μ	parameter of unscaled Cohen model, see equation (3.1), page 33
$\boldsymbol{n}$	unit outward normal vector of Ω, page 17
ν	parameter of unscaled Cohen model, see equation (3.1), page 33
Ω	connected, open, bounded domain in $\mathbb{R}^d$, page 17
$\bar{\Omega}$	closure of Ω, i.e. $\Omega \cup \partial\Omega$
$\partial\Omega$	boundary of domain Ω, page 17
σ	'stress' variable in Cohen model, see equation (3.1), page 33
$\breve{\sigma}$	initial condition on σ, see equation (3.1), page 33
σ_D	$\in \{0, 1\}$, indicates whether Dirichlet boundary conditions for σ are prescribed, see equation (7.2), page 101
T	endpoint of time interval, page 22
t	time variable, page 22
u	solvent concentration in Cohen model, see equation (3.1), page 33
u_a	midpoint of transition region of γ, page 37
$\breve{u}$	initial condition on u, see equation (3.1), page 33
u_D	inhomogeneous Dirichlet boundary conditions, see equation (3.1), page 33

Finite Element Notation

$\boldsymbol{a}_j$	nodes of Lagrange basis functions, see equation (2.28), page 27
$\mathcal{E}_h$	the mesh, page 26
$\mathcal{G}^h$	discrete inverse Laplacian, see equation (5.12), page 62
h	upper bound for element diameters of mesh $\mathcal{E}_h$, page 26
h_E	diameter of element E, page 26
$\hat{h}$	upper bound for h, page 26
I^h	nodal interpolation operator, see equation (2.31), page 28
N_V	number of Lagrange basis functions in V^h, dimension of V^h, see equation (2.28), page 27
N_W	number of Lagrange basis functions in W^h, dimension of W^h, see equation (2.28), page 27
ϕ_j	Lagrange basis function, see equation (2.29), page 27
Π_0^h	L_2-projection, page 30
Π_V^h	elliptic projection, page 30
r	polynomial degree of Lagrange basis functions in V^h, see equation (2.27), page 27
ρ_E	diameter of the largest ball that can be inscribed in element E, page 26
$\mathcal{SZ}^h$	Scott-Zhang interpolation operator, page 30
V^h	finite element space with homogeneous Dirichlet boundary conditions, see equation (2.27), page 27
W^h	finite element space, see equation (2.27), page 27

Time discretisation

k	constant length of time steps, see equation (5.1), page 59
N	number of time steps, page 59

t_n	points of time discretisation, page 59
v_n	shorthand notation for $v(t_n)$, see equation (5.2), page 60
$\bar{v}_n$	$\frac{1}{2}(v_n + v_{n-1})$, see equation (5.3), page 60
$\partial_t v_n$	$\frac{1}{k}(v_n - v_{n-1})$, see equation (5.3), page 60
$\partial_t^2 v_n$	$\frac{1}{k^2}(v_n - 2v_{n-1} + v_{n-2})$, see equation (5.3), page 60
Δ_n	error in the trapezium rule, see equation (5.3), page 60

Discretised Problem

σ_n^h	discrete approximation to $\sigma(t_n)$, see equation (5.15), page 62
u_n^h	discrete approximation to $u(t_n)$, see equation (5.15), page 62
σ_n^*	elliptic projection of σ_n, see equation (6.1), page 79
u_n^*	elliptic projection of u_n, see equation (6.1), page 79
χ_n	$= u_n^h - u_n^*$, see equation (6.2), page 79
η_n	$= \sigma_n^h - \sigma_n^*$, see equation (6.2), page 79
θ_n	$= \sigma_n - \sigma_n^*$, see equation (6.2), page 79
ξ_n	$= u_n - u_n^*$, see equation (6.2), page 79
Q	index for numerical method $Q \in \{1, 2, 3, 4, 5\}$, page 63
$\mathscr{B}_n^Q$	discrete approximation of $\overline{\gamma(u)\sigma}_n$, page 63
$\mathscr{E}_n^Q$	extrapolation operator for u^h, page 63
E_n^Q	auxiliary term containing the extrapolation error, see equation (6.39), page 89
$\hat{k}$	maximum time step size for $Q = 4, 5$ in error estimate, page 94

Matrix formulation

$\underline{\boldsymbol{A}}$	stiffness matrix, see equation (5.19), page 64
$\underline{\boldsymbol{B}}^n$	a diagonal matrix, see equation (5.33), page 66
$\underline{\boldsymbol{C}}$	mass matrix, see equation (5.19), page 64
$\underline{\boldsymbol{D}}^n$	a diagonal matrix, see equation (5.33), page 66
$\underline{\boldsymbol{\Gamma}}^n$	a diagonal matrix, see equation (5.20), page 64
$\underline{\boldsymbol{I}}$	identity matrix
$\boldsymbol{L}^n$	load vector for u-equation, see equation (5.21), page 64
$\hat{\boldsymbol{L}}^n$	load vector for σ-equation, see equation (5.21), page 64
$\boldsymbol{S}^n$	solution vector for σ_n^h, see equation (5.22), page 64
$\boldsymbol{U}^n$	solution vector for u_n^h, see equation (5.22), page 64
$\underline{\boldsymbol{\Upsilon}}^n$	a diagonal matrix, see equation (5.20), page 64

Convergence Tables

e_{comb}	combined error for convergence tests, see equation (7.5), page 103
e_1	H^1-part of error for convergence tests, see equation (7.3), page 103
e_0	L_2-part of error for convergence tests, see equation (7.2), page 103
EOC	estimated order of convergence, see equation (7.6), page 105
ndof	number of degrees of freedom, page 105

Behaviour of Solutions

Δ^*	critical value of Δ, page 146

$D(u)$	auxiliary function, nonlinear diffusion coefficient, see equation (7.9), page 146
$G(u)$	auxiliary function, see equation (7.9), page 146
$M_u(t)$	mass uptake of u, see equation (7.7), page 144
u^*	critical value of u for Type II, page 147
u_+	critical value of u for Type III, page 147
u_-	critical value of u for Type III, page 147
u_c	critical value of u for Type III, page 147
u_s	Dirichlet boundary value for u, page 146
$x_m(t)$	position of largest gradient of u, page 144

Index

A
Aubin-Nitsche Lemma, 32

B
Banach's fixed point theorem, 177
basis functions, 27

C
Cauchy-Schwarz inequality, 19
CG, 102
Cohen model, 33
colour map, 160
concave function, 177
continuous embedding, 181
contraction, 177
convex function, 177
Crank-Nicolson method, 62

D
dense embedding, 181
Dirichlet boundary, 17
discretisation
    of nonlinearity, 63
    space, 26
    time, 59
distributional derivatives, 18
domain, 17
dual space, 180
duality pairing, 180

E
elliptic regularity, 17
estimated order of convergence, 105
exactness, 103

F
finite element space, 27

G
Gelfand triple, 183
GMRes, 102
Grönwall's lemma
    continuous, 175
    discrete, 176

H
Hahn-Banach theorem, 181

I
interpolation
    nodal, 28
    Scott-Zhang, 30
inverse estimates, 28
inverse Laplacian, 22
    discrete, 62

J
Jensen's inequality, 175

L
Lagrange nodes, 27
    global, 27
    local, 27
Lebesgue spaces, 18
load vector, 64

M
mass matrix, 64
mass uptake, 144
mesh, 26

N
Neumann boundary, 17
notation, 16

P
Poincaré's inequality, 20
polymer
    glassy, 35
    rubbery, 35
positive definite matrix, 183

R
reflexive space, 182
Riesz representation theorem, 182

S
Scott-Zhang interpolation, 30
sharp front, 144
simplex, 26
Sobolev spaces, 18
sorption overshoot, 144
space discretisation, 26
stiffness matrix, 64

T
time discretisation, 59

U
unit outward normal vector, 17

V
viscoelastic, 34

W
weighted norm, 48

Y
Young's inequality, 175