

# Discontinuous Galerkin finite element approximation of nonlinear non-Fickian diffusion in viscoelastic polymers

Béatrice Rivière\* and Simon Shaw†

May 4, 2006

## Abstract

We consider discrete schemes for a nonlinear model of non-Fickian diffusion in viscoelastic polymers. The model is motivated by, but not the same as, that proposed by Cohen *et al.* in *SIAM J. Appl. Math.*, **55**, pp. 348–368, 1995. The spatial discretisation is effected with both the symmetric and non-symmetric interior penalty discontinuous Galerkin finite element method, and the time discretisation is of Crank-Nicolson type. We also discuss two means of handling the nonlinearity: either implicitly, which requires the solution of nonlinear equations at each time level, or through a linearisation based on extrapolating from previous time levels. The same optimal orders of convergence are proven in both cases and, to verify this, some numerical results are also given for the linearised scheme.

## 1 Introduction

In [25] Thomas and Windle demonstrated by experiment that the diffusion of organic penetrants into glassy polymers does not obey the classical Fick’s law. At moderate temperatures the profile of diffusing penetrant (methanol in their case) forms a steep front which travels at a constant speed into the polymer. In [24] they developed a model for this ‘anomalous’ diffusion in terms of an ordinary differential equation for the fractional swelling of the polymer.

However, in order to have more predictive value, a mathematical model for this behaviour in the form of a partial differential equation is more desirable. Such a model has been proposed by Cohen *et al.* in [6] (see also the references therein). Recognising that viscoelastic stress relaxation effects are significant in polymers, they add such a term to Fick’s law, and drive this stress through a nonlinear relaxation equation which is adjoined to the diffusion equation. Solving the system then results in a heat equation with a nonlinear viscoelastic memory term in the form of a Volterra integral—typical of continuum models of polymers (see e.g. [8] for polymer theory and [14] for a similar model of heat conduction).

In terms of the underlying physics, it seems that high levels of penetrant concentration can cause a *rubber-glass phase change*. The polymer’s viscoelastic properties change dramatically across this transition layer, and this can cause sharp fronts to develop in the diffusing penetrant.

The model proposed in [6] seems to be difficult to handle in terms of obtaining estimates and so, as a stepping stone to that model, we deal here with a simpler version which involves a

---

\*Computational Mathematics Research Group, Department of Mathematics, 301 Thackeray, University of Pittsburgh, Pittsburgh, PA 15260. (riviere@math.pitt.edu).

†BICOM (Brunel Institute of Computational Mathematics), Brunel University, Uxbridge UB8 3PH, England. (simon.shaw@brunel.ac.uk). Shaw would like to acknowledge the support of the Engineering and Physical Sciences Research Council, GR/R10844/01, and also the US Army Research Office, DAAD19-00-1-0421.

vector of stresses in the diffusion equation, rather than (as in [6]) the gradient of a scalar stress. The reason why this is a simplification can be better explained once we have seen the equations.

Our model is as follows. For an open bounded domain  $\Omega \subset \mathbb{R}^d$  ( $d = 2$  or  $3$ ) and a time interval  $I := (0, T)$ , for some  $T > 0$ , we want to find the ‘concentration’,  $u: \Omega \times I \rightarrow \mathbb{R}$ , and viscoelastic stress,  $\boldsymbol{\sigma}: \Omega \times I \rightarrow \mathbb{R}^d$ , such that in  $\Omega \times I$ ,

$$u_t(t) - \nabla \cdot D \nabla u(t) = f(t) + \nabla \cdot K \boldsymbol{\sigma}(t), \quad (1)$$

$$\boldsymbol{\sigma}_t(t) + \gamma(u) \boldsymbol{\sigma}(t) = \mu \nabla u(t), \quad (2)$$

where  $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_d)^T$ . These are subject to the initial conditions,

$$u(\mathbf{x}, 0) = \check{u}(\mathbf{x}) \quad \text{and} \quad \boldsymbol{\sigma}(\mathbf{x}, 0) = \check{\boldsymbol{\sigma}}(\mathbf{x}), \quad (3)$$

and the boundary conditions,

$$u(\mathbf{x}, t) = 0 \text{ on } \Gamma_D \times I \quad \text{and} \quad (D \nabla u(\mathbf{x}, t) + K \boldsymbol{\sigma}(\mathbf{x}, t)) \cdot \mathbf{n}(\mathbf{x}) = g(\mathbf{x}, t) \text{ on } \Gamma_N \times I, \quad (4)$$

where  $\Gamma_D \cup \Gamma_N = \partial\Omega$ ,  $\Gamma_D \cap \Gamma_N = \emptyset$ ,  $\Gamma_N$  has outward normal  $\mathbf{n}$  and  $\Gamma_D$  is closed with positive surface measure. Note that in (1) and (2), and usually below, we drop the  $\mathbf{x}$  dependence.

In [6] the vector of stresses,  $\boldsymbol{\sigma}$ , is replaced by the gradient of a scalar stress,  $\nabla \sigma$ . Our model is a simplification because, in weak form, we can generate the term  $(\boldsymbol{\sigma}, \nabla u)$  in both equations and therefore easily merge them as a starting point for estimates. This is clearly illustrated below in Theorem 1.1.

In our equations  $D$ ,  $K$  and  $\mu$  are positive constants. Also, the nonlinear function

$$\gamma(u) = \frac{1}{2}(\gamma_R + \gamma_G) + \frac{1}{2}(\gamma_R - \gamma_G) \tanh\left(\frac{u - u_{RG}}{\Delta}\right), \quad (5)$$

with constants  $\gamma_R \gg \gamma_G > 0$ , models the sharp change in material properties across the *rubber-glass transition*. The sharpness of the change is controlled by the positive constant  $\Delta$ , and the location of the change is controlled by the constant transition concentration  $u_{RG}$ . Regions where  $u \ll u_{RG}$  correspond to the ‘glassy’ phase while regions where  $u \gg u_{RG}$  are ‘rubbery’. When  $u$  is in or near a  $\Delta$ -neighbourhood of  $u_{RG}$  the polymer is in a nebulous phase transition state.

Since this simplified model is motivated by a diffusion problem we have continued to refer to  $u$  as a concentration. However, because the underlying physics may have been lost in the simplification, it will not actually have the correct physical properties. For example, the computed solutions shown later in Figure 1 clearly show  $u > 1$  which, if interpreted physically, suggest a concentration of greater than 100%.

We note that,

$$0 < \gamma_G \leq \gamma(y) \leq \gamma_R \quad \forall y \in \mathbb{R}, \quad (6)$$

and,

$$\gamma'(y) = \frac{\gamma_R - \gamma_G}{2\Delta} \operatorname{sech}^2\left(\frac{y - u_{RG}}{\Delta}\right), \quad (7)$$

so that,

$$0 \leq \gamma'(y) \leq C'_\gamma := \frac{\gamma_R - \gamma_G}{2\Delta} \quad \forall y \in \mathbb{R}. \quad (8)$$

Also,

$$\gamma''(y) = -\left(\frac{\gamma_R - \gamma_G}{\Delta^2}\right) \tanh\left(\frac{y - u_{RG}}{\Delta}\right) \operatorname{sech}^2\left(\frac{y - u_{RG}}{\Delta}\right),$$

which gives,

$$|\gamma''(y)| \leq C''_\gamma := \frac{\gamma_R - \gamma_G}{\Delta^2} \quad \forall y \in \mathbb{R}. \quad (9)$$

We also note that we can solve (2) to get,

$$\boldsymbol{\sigma}(t) = \check{\boldsymbol{\sigma}} e^{-\int_0^t \gamma(u(\xi)) d\xi} + \mu \int_0^t e^{-\int_s^t \gamma(u(\xi)) d\xi} \nabla u(s) ds, \quad (10)$$

and use this in (1) to arrive at (assuming  $\check{\boldsymbol{\sigma}} = \mathbf{0}$ ),

$$u_t(t) - \nabla \cdot D \nabla u(t) = f(t) + \nabla \cdot \mu K \int_0^t e^{-\int_s^t \gamma(u(\xi)) d\xi} \nabla u(s) ds. \quad (11)$$

We recognise this as a parabolic partial differential equation with a nonlinear Volterra-type memory term typical of that arising in viscoelasticity theory. We could work directly with this formulation in constructing our numerical approximation, but we prefer to work with the system, (1) with (2), since we then need not be concerned with the discretisation of the Volterra integral. Also, representing viscoelasticity through evolution equations for internal variables is often preferred to the use of Volterra integrals. See for example [11, 10, 3]. It is important to realise that introducing internal variables does not introduce more unknowns and lead to a more complex scheme than would result from using the Volterra formulation directly. In the latter case the ‘history’ in the Volterra integral needs to be stored and updated at each time step. This is exactly analogous to storing the previous value of the internal variable and then updating it through a time-stepping scheme.

This is the third in a series of papers extending the (spatially) *discontinuous Galerkin finite element method* (DG FEM) to viscoelasticity problems. In [19] we considered an elliptic stress analysis problem with memory and in [20] we extended this to a second-order hyperbolic problem with memory. Both of these deal only with linear problems but below we ‘complete the set’ by considering a parabolic problem, and including a physically relevant nonlinearity.

Discontinuous Galerkin methods offer several advantages. The lack of continuity constraints between the local approximations allows for an easy implementation of mesh adaptivity. Unlike the classical continuous finite element method, the DG method can handle unstructured nonconforming meshes with several hanging nodes per edge (or face). In addition, increasing the polynomial degree does not require any major modification of the software. It is relatively easy as well to have polynomial degrees that vary from one mesh element to the next. Finally, one inherent property to DG methods is the local mass conservation. While this property is essential in many flow and transport problems, it remains to be seen that local mass conservation is important for non-Fickian polymer diffusion problems. A deeper numerical investigation is needed. This will be the object of future work.

The layout of this article is as follows. In Section 2 the equations are spatially discretised using an interior penalty DG FEM, and we consider both the symmetric and non-symmetric variants. The time discretisation is a standard Crank-Nicolson method with a choice of treatments for the nonlinear term. Either this term is approximated in an implicit way, which involves a nonlinear equation set at each time level, or it is handled by extrapolating the current approximation of  $u$  to the current time level from the two previous time levels (similarly to [5]). Special care is needed at the first time step, but we can show optimal second-order convergence in each case. The error estimates are contained in Section 3 and some numerical experiments are given in Section 4. We finish with some comments regarding our model and approach in Section 5, as well as discuss the potential for extending this work to Cohen *et al.*’s model.

For background to the DG FEM we refer to Rivière *et al.* in [16, 15, 9, 21], and for the numerical analysis of generic parabolic problems with memory we refer to [5, 12, 23, 26, 13] (but there are many others).

However, apart from [20], we are not aware of any error analysis for numerical approximations to viscoelasticity problems where the Volterra integral is replaced with internal variable evolution equations, such as (2).

Our notation is standard. For  $\omega \subseteq \bar{\Omega}$  we use  $(\cdot, \cdot)_\omega$  to denote the  $L_2(\omega)$  inner product and simply write  $(\cdot, \cdot)$  when  $\omega = \Omega$ . Also, we use  $\|\cdot\|_{p,\omega}$  to denote the  $H^p(\omega)$  norm and write  $\|\cdot\|_m$  as an abbreviation for  $\|\cdot\|_{m,\Omega}$ .

We set  $\mathbf{H}^p(\Omega) := (H^p(\Omega))^d$ , but in the notation just described we do not distinguish between inner products and norms on  $H^p(\Omega)$  (as used for  $u$ ) and inner products and norms on  $\mathbf{H}^p(\Omega)$  (as used for  $\boldsymbol{\sigma}$ ).

Since our functions are time dependent we take the usual approach of thinking of them as maps from  $I$  to some underlying Banach space,  $X$ . For  $1 \leq p < \infty$  the  $L_p(0, t; X)$  norms are given by,

$$\|v\|_{L_p(0,t;X)} := \left( \int_0^t \|v(t)\|_X^p dt \right)^{1/p},$$

with the usual 'ess sup' modification when  $p = \infty$ .

To finish this introduction we derive a stability estimate for this problem, the proof is a model of how to proceed with the estimates for the discrete scheme that follows. To begin we note that if,

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

then a variational formulation of (1)-(2) is: find maps  $u : I \rightarrow V$  and  $\boldsymbol{\sigma} : I \rightarrow \mathbf{L}_2(\Omega)$  such that

$$(u_t(t), v) + (D\nabla u(t), \nabla v) + (K\boldsymbol{\sigma}(t), \nabla v) = L(t; v) \quad \forall v \in V, \quad (12)$$

$$(\boldsymbol{\sigma}_t(t) + \gamma(u)\boldsymbol{\sigma}(t), \mathbf{w}) = (\mu\nabla u(t), \mathbf{w}) \quad \forall \mathbf{w} \in \mathbf{L}_2(\Omega), \quad (13)$$

where,

$$L(t; v) := (f(t), v) + (g(t), v)_{\Gamma_N}. \quad (14)$$

We can now state a basic stability estimate which does not require Gronwall's lemma.

**Theorem 1.1 (basic stability)** *There exists a constant  $C > 0$ , independent of  $T$ , such that, if  $(u, \boldsymbol{\sigma})$  is a solution of (12), (13), then*

$$\begin{aligned} & \|u(t)\|_0^2 + \|\boldsymbol{\sigma}(t)\|_0^2 + \int_0^t \left( \|D^{1/2}\nabla u(s)\|_0^2 + \|\boldsymbol{\sigma}(s)\|_0^2 \right) ds \\ & \leq C \left( \|\ddot{u}\|_0^2 + \|\ddot{\boldsymbol{\sigma}}\|_0^2 + \|f\|_{L_2(0,t;L_2(\Omega))}^2 + \|g\|_{L_2(0,t;L_2(\Gamma_N))}^2 \right) \end{aligned}$$

for all  $t > 0$ .

**Proof** Choose  $v = u$  in (12) and  $\mathbf{w} = (K/\mu)\boldsymbol{\sigma}(t)$  in (13) and add the resulting equations to get,

$$\begin{aligned} & (u_t(t), u(t)) + (D\nabla u(t), \nabla u(t)) + (K\boldsymbol{\sigma}(t), \nabla u(t)) \\ & + \frac{K}{\mu}(\boldsymbol{\sigma}_t(t), \boldsymbol{\sigma}(t)) + \frac{K}{\mu}(\gamma(u)\boldsymbol{\sigma}(t), \boldsymbol{\sigma}(t)) - (K\boldsymbol{\sigma}(t), \nabla u(t)) \\ & = (f(t), u(t)) + (g(t), u(t))_{\Gamma_N}. \end{aligned}$$

Hence, using Poincaré's inequality

$$\begin{aligned} & \frac{d}{dt}\|u(t)\|_0^2 + \frac{K}{\mu} \frac{d}{dt}\|\boldsymbol{\sigma}(t)\|_0^2 + 2\|D^{1/2}\nabla u(t)\|_0^2 + \frac{2K}{\mu}(\gamma(u)\boldsymbol{\sigma}(t), \boldsymbol{\sigma}(t)) \\ & \leq 2C\|f(t)\|_0\|D^{1/2}\nabla u(t)\|_0 + 2C\|g(t)\|_{0,\Gamma_N}\|D^{1/2}\nabla u(t)\|_0 \\ & \leq 2C^2\|f(t)\|_0^2 + 2C^2\|g(t)\|_{0,\Gamma_N}^2 + \|D^{1/2}\nabla u(t)\|_0^2. \end{aligned}$$

Integrating then gives,

$$\begin{aligned} \|u(t)\|_0^2 + \frac{K}{\mu} \|\boldsymbol{\sigma}(t)\|_0^2 + \int_0^t \left( \|D^{1/2} \nabla u(s)\|_0^2 + \frac{2K\gamma_G}{\mu} \|\boldsymbol{\sigma}(s)\|_0^2 \right) ds \\ \leq \|\check{u}\|_0^2 + \frac{K}{\mu} \|\check{\boldsymbol{\sigma}}\|_0^2 + 2C^2 \left( \|f\|_{L_2(0,t;L_2(\Omega))}^2 + \|g\|_{L_2(0,t;L_2(\Gamma_N))}^2 \right). \end{aligned}$$

This concludes the proof.  $\square$

Lastly in this section, we recall Young's inequality in the form,

$$ab \leq \frac{a^p}{p\epsilon^p} + \frac{\epsilon^q b^q}{q}, \quad (15)$$

for all  $a, b \geq 0$ ,  $\epsilon > 0$  and  $p, q \in (1, \infty)$  such that  $1/p + 1/q = 1$ .

## 2 The numerical scheme

The first step is to establish notation for the spatial discretisation. Let  $\mathcal{E}_h = \{E\}$  be a nondegenerate quasiuniform subdivision of  $\Omega$ , where  $E$  is a triangle if  $d = 2$ , or a tetrahedron if  $d = 3$ . The nondegeneracy requirement is that there exists  $\rho > 0$  such that if  $h_E = \text{diam}(E)$ , then  $E$  contains a ball of radius  $\rho h_E$  in its interior. Let  $h = \max\{h_E : E \in \mathcal{E}_h\}$ , the quasiuniformity requirement is that there exists  $\tau > 0$  such that  $h/h_E \leq \tau$  for all  $E \in \mathcal{E}_h$ . We denote by  $\Gamma_h$  the set of interior edges (faces for  $d = 3$ ) of  $\mathcal{E}_h$ . With each edge (or face)  $e$ , we associate a unit normal vector  $\mathbf{n}_e$ . For a boundary edge  $e$ ,  $\mathbf{n}_e$  is taken to be the unit outward vector normal to  $\partial\Omega$ .

We now define the average and the jump operators. For each of the interior edges, suppose that  $e$  is shared by  $E_1^e$  and  $E_2^e$  such that  $\mathbf{n}_e$  points from  $E_1^e$  to  $E_2^e$  and for a boundary edge, suppose that  $e$  belongs to  $E_1^e$ . We define the averaging operator  $\{\cdot\}$  by,

$$\{w\} := \begin{cases} \frac{1}{2}(w|_{E_1^e})|_e + \frac{1}{2}(w|_{E_2^e})|_e & \text{if } e \subset \Omega, \\ (w|_{E_1^e})|_e & \text{if } e \subset \partial\Omega. \end{cases}$$

and the jump operator  $[\cdot]$  by,

$$[w] := \begin{cases} (w|_{E_1^e})|_e - (w|_{E_2^e})|_e & \text{if } e \subset \Omega, \\ (w|_{E_1^e})|_e & \text{if } e \subset \partial\Omega. \end{cases}$$

The distinction between  $[\cdot]$  and  $-[\cdot]$  can be made because each edge  $e_a$  has a unit normal associated with it. The "direction" in which the jump takes place is unimportant.

These operators are well defined if  $w|_{E_a^i} \in H^{\frac{1}{2}+\epsilon}(E_a^i)$  for  $i = 1, 2$  and  $\epsilon > 0$ . Below, we use  $|e|$  to denote the  $(d-1)$ -dimensional surface measure of the edge/face  $e$ . We also frequently use the estimate,  $|e| \leq Ch^{d-1}$  which arises as a consequence of our assumptions.

Define the broken spaces for any integer  $r > 0$ ,

$$\begin{aligned} \mathcal{D}_r(\mathcal{E}_h) &= \{v \in L_2(\Omega) : v|_E \in \mathbb{P}_r(E) \quad \forall E \in \mathcal{E}_h\}, \\ \mathcal{D}_r(\mathcal{E}_h) &= \mathcal{D}_r(\mathcal{E}_h)^d, \\ H^n(\mathcal{E}_h) &= \{v \in L_2(\Omega) : v|_E \in H^n(E) \quad \forall E \in \mathcal{E}_h\}. \end{aligned}$$

For these finite element spaces we have the following interpolation-error estimates. If  $v \in H^n(\mathcal{E}_h) \cap C(\bar{\Omega})$  and  $\mu = \min\{r+1, n\}$  then there is an interpolant  $\hat{v} \in \mathcal{D}_r(\mathcal{E}_h) \cap C(\bar{\Omega})$  such that

for each  $E \in \mathcal{E}_h$ ,

$$\|v - \hat{v}\|_{m,E} \leq Ch_E^{\mu-m} \|v\|_{n,E} \quad \text{for } n \geq m \geq 0, \quad (16)$$

$$\|v - \hat{v}\|_{m,\gamma} \leq Ch_E^{\mu-m-1/2} \|v\|_{n,E} \quad \text{for } m = 0, 1 \text{ and } n \geq m, \quad (17)$$

where  $\gamma \subseteq \partial E$ .

Define the bilinear forms

$$\begin{aligned} J_0^{\delta,\beta}(w, v) &= \sum_{e \in \Gamma_h \cup \Gamma_D} \frac{\delta}{|e|^\beta} \int_e [w][v] \quad \text{for } \beta \geq (d-1)^{-1}, \\ A(w, v) &= \sum_E \int_E D \nabla w \cdot \nabla v + J_0^{\delta,\beta}(w, v) - \sum_{e \in \Gamma_h \cup \Gamma_D} \int_e \{D \nabla w \cdot \mathbf{n}_e\} [v] \\ &\quad + \kappa \sum_{e \in \Gamma_h \cup \Gamma_D} \int_e \{D \nabla v \cdot \mathbf{n}_e\} [w]. \end{aligned}$$

Here  $\kappa$  is a switch: we set  $\kappa = 1$  to obtain the non-symmetric DG scheme, and  $\kappa = -1$  to obtain the symmetric scheme. Following from these definitions are the norm and semi-norm,

$$\|v\|_{\mathcal{A}} := \left( |v|_{\mathcal{E}}^2 + J_0^{\delta,\beta}(v, v) \right)^{\frac{1}{2}} \quad \text{and} \quad |v|_{\mathcal{E}} := \left( \sum_{E \in \mathcal{E}_h} \int_E D \nabla v \cdot \nabla v \, dE \right)^{\frac{1}{2}}.$$

We will need the following estimates.

**Lemma 2.1** *We have,*

$$\|v\|_0 \leq C_f \|v\|_{\mathcal{A}} \quad \forall v \in H^1(\mathcal{E}_h), \quad (18)$$

and

$$\|v\|_{0,\Gamma_N} \leq C_g h^{-1/2} \|v\|_{\mathcal{A}} \quad \forall v \in \mathcal{D}_r(\mathcal{E}_h),$$

for constants  $C_f$  and  $C_g$ , independent of  $h$ .

**Proof.** For the first inequality we refer to [9, Lemma 6.2], and for the second inequality we use the first one with Sobolev interpolation to get,

$$\|v\|_{0,\Gamma_N}^2 = \sum_{e \in \Gamma_N} \|v\|_{0,e}^2 \leq C \sum_E h^{-1} \|v\|_{0,E} \|\nabla v\|_{0,E} \leq Ch^{-1} (C_f^2 + D^{-1}) \|v\|_{\mathcal{A}}^2.$$

This completes the proof.  $\square$

We note that if  $u(t) \in C(\bar{\Omega})$  for each  $t$  then,

$$\begin{aligned} (u_t(t), v) + A(u(t), v) &= L(t; v) - \sum_E (K \boldsymbol{\sigma}(t), \nabla v)_E \\ &\quad + \sum_{e \in \Gamma_h \cup \Gamma_D} \int_e \{K \boldsymbol{\sigma}(t) \cdot \mathbf{n}_e\} [v] \quad \forall v \in \mathcal{D}_r(\mathcal{E}_h), \end{aligned} \quad (19)$$

and

$$(\boldsymbol{\sigma}_t(t) + \gamma(u) \boldsymbol{\sigma}(t), \mathbf{w}) = \sum_E (\mu \nabla u(t), \mathbf{w})_E \quad \forall \mathbf{w} \in \mathcal{D}_{r-1}(\mathcal{E}_h). \quad (20)$$

The first of these arises by element-wise partial integration and ‘adding zero’ (see e.g. [15]).

To construct a fully discrete approximation we set  $k = T/N$ , for some  $N \in \mathbb{N}$ , and write  $t_i = ik$ . To ease notation we define,

$$\partial_t w_i := \frac{w(t_i) - w(t_{i-1})}{k} \quad \text{and} \quad \bar{w}_i := \frac{w(t_i) + w(t_{i-1})}{2}.$$

The fully discrete approximations,  $u^h$  and  $\sigma^h$ , to  $u$  and  $\sigma$  are continuous and piecewise linear in time, and discontinuous in space. We set  $u_i^h := u^h(t_i)$  and  $\sigma_i^h := \sigma^h(t_i)$ .

An issue is how to handle the nonlinearity,  $\gamma(u)$ , in the numerical scheme. We offer two possibilities by approximating  $\gamma(u)|_{(t_{i-1}, t_i)}$  by  $\gamma(\mathcal{B}_{i,n}u^h)$ , for  $n = 1$  or  $2$ , where,

$$\mathcal{B}_{i,1}u^h := \bar{u}_i^h \quad \text{and} \quad \mathcal{B}_{i,2}u^h := \mathcal{E}_i u^h,$$

with  $\mathcal{E}_i$  an extrapolation operator defined by,

$$\mathcal{E}_i u^h := \begin{cases} u_0^h & \text{for } i = 1; \\ \frac{3}{2}u_{i-1}^h - \frac{1}{2}u_{i-2}^h & \text{for } i = 2, \dots, N. \end{cases}$$

In the first case we approximate  $\gamma(u)|_{(t_{i-1}, t_i)}$  by taking the true average,  $\bar{u}_i^h$ , of the discrete solution. This will result in a nonlinear system to be solved at each time level. To linearise this system, the second method linearly extrapolates to the average based on the two previous solutions. This is not possible at the first time step and so this first step will require special treatment in the error estimation. This extrapolation technique is widely used in coupled flow and transport problems such as miscible displacement. See, for example, [7] and [17].

The fully discrete approximations (i.e. for  $n = 1$  or  $2$ ) are based on sampling (19) and (20) at the temporal midpoints,  $t_{i-1/2}$ . They are defined to be: for each  $i = 1, 2, \dots, N$ , find a pair  $\{u_i^h, \sigma_i^h\} \in \mathcal{D}_r(\mathcal{E}_h) \times \mathcal{D}_{r-1}(\mathcal{E}_h)$  such that,

$$\begin{aligned} (\partial_t u_i^h, v) + A(\bar{u}_i^h, v) &= L_i(v) - \sum_E (K \bar{\sigma}_i^h, \nabla v)_E \\ &+ \sum_{e \in \Gamma_h \cup \Gamma_D} \int_e \{K \bar{\sigma}_i^h \cdot \mathbf{n}_e\} [v] \quad \forall v \in \mathcal{D}_r(\mathcal{E}_h), \end{aligned} \quad (21)$$

and

$$(\partial_t \sigma_i^h + \gamma(\mathcal{B}_{i,n}u^h) \bar{\sigma}_i^h, \mathbf{w}) = \sum_E (\mu \nabla \bar{u}_i^h, \mathbf{w})_E \quad \forall \mathbf{w} \in \mathcal{D}_{r-1}(\mathcal{E}_h), \quad (22)$$

where,

$$L_i(v) := \frac{1}{2} (L(t_i; v) + L(t_{i-1}; v)),$$

and the discrete initial data are given by,

$$\begin{aligned} (u_0^h, v) &= (\check{u}, v) & \forall v \in \mathcal{D}_r(\mathcal{E}_h), \\ (\sigma_0^h, \mathbf{w}) &= (\check{\sigma}, \mathbf{w}) & \forall \mathbf{w} \in \mathcal{D}_{r-1}(\mathcal{E}_h). \end{aligned}$$

We now give a stability estimate for this discrete approximation and note that Gronwall's lemma is not used. We also note that the ' $h^{-1}$ ' factor appearing in front of the boundary term is a weakness in the proof and is not observed in computations. It appears that the removal of this factor is an open problem (although see Remark 3.6 later).

**Theorem 2.2 (discrete basic stability)** *If  $\beta \geq (d-1)^{-1}$  and  $h \leq \hat{h}$  we have for  $m = 1, 2, \dots, N$  that,*

$$\begin{aligned} \|u_m^h\|_0^2 + \frac{K}{\mu} \|\sigma_m^h\|_0^2 + C^* k \sum_{i=1}^m \left( \|\bar{u}_i^h\|_{\mathcal{A}}^2 + 2K \|\bar{\sigma}_i^h\|_0^2 \right) \\ \leq \|\check{u}\|_0^2 + \frac{K}{\mu} \|\check{\sigma}\|_0^2 + 6k \sum_{i=1}^m (C_f^2 \|\bar{f}_i\|_0^2 + C_g^2 h^{-1} \|\bar{g}_i\|_{0, \Gamma_N}^2), \end{aligned}$$

provided that,

$$\delta \geq 3C\hat{h}^{(d-1)\beta-1} \max \left\{ \frac{4D}{1-C^*}, \frac{\mu K}{2\gamma_G - 2\mu C^*} \right\},$$

where:  $C^* < \min\{1, \gamma_G/\mu\}$  is some chosen positive constant;  $C$  is independent of  $h$ ; and,  $C_f$  and  $C_g$  are those in Lemma 2.1.

**Proof.** Choose  $v = \bar{u}_i^h$  in (21) and  $w = (K/\mu)\bar{\sigma}_i^h$  in (22) and note that,

$$\begin{aligned} (\partial_t u_i^h, \bar{u}_i^h) &= \frac{1}{2k} \|u_i^h\|_0^2 - \frac{1}{2k} \|u_{i-1}^h\|_0^2, \\ (\partial_t \sigma_i^h, \bar{\sigma}_i^h) &= \frac{1}{2k} \|\sigma_i^h\|_0^2 - \frac{1}{2k} \|\sigma_{i-1}^h\|_0^2, \\ A(\bar{u}_i^h, \bar{u}_i^h) &= \sum_E (D\nabla \bar{u}_i^h, \nabla \bar{u}_i^h)_E + J_0^{\delta, \beta}(\bar{u}_i^h, \bar{u}_i^h) \\ &\quad + (\kappa - 1) \sum_{e \in \Gamma_h \cup \Gamma_D} \int_e \{D\nabla \bar{u}_i^h \cdot \mathbf{n}_e\} [\bar{u}_i^h]. \end{aligned}$$

Adding the two resulting equations then gives,

$$\begin{aligned} &\frac{1}{2k} \|u_i^h\|_0^2 - \frac{1}{2k} \|u_{i-1}^h\|_0^2 + \frac{K}{2k\mu} \|\sigma_i^h\|_0^2 - \frac{K}{2k\mu} \|\sigma_{i-1}^h\|_0^2 + \|\bar{u}_i^h\|_{\mathcal{A}}^2 + \frac{K}{\mu} (\gamma(\mathcal{B}_{i,n} u^h) \bar{\sigma}_i^h, \bar{\sigma}_i^h) \\ &= L_i(\bar{u}_i^h) - (\kappa - 1) \sum_{e \in \Gamma_h \cup \Gamma_D} \int_e \{D\nabla \bar{u}_i^h \cdot \mathbf{n}_e\} [\bar{u}_i^h] + \sum_{e \in \Gamma_h \cup \Gamma_D} \int_e \{K\bar{\sigma}_i^h \cdot \mathbf{n}_e\} [\bar{u}_i^h], \end{aligned}$$

and summing over  $i = 1, \dots, m$  and multiplying by  $2k$  yields,

$$\begin{aligned} &\|u_m^h\|_0^2 + \frac{K}{\mu} \|\sigma_m^h\|_0^2 + 2k \sum_{i=1}^m \|\bar{u}_i^h\|_{\mathcal{A}}^2 + 2k \sum_{i=1}^m \frac{K}{\mu} (\gamma(\mathcal{B}_{i,n} u^h) \bar{\sigma}_i^h, \bar{\sigma}_i^h) \\ &= \|u_0^h\|_0^2 + \frac{K}{\mu} \|\sigma_0^h\|_0^2 + 2k \sum_{i=1}^m L_i(\bar{u}_i^h) + I + II, \end{aligned}$$

where,

$$\begin{aligned} I &= 2k \sum_{i=1}^m \sum_{e \in \Gamma_h \cup \Gamma_D} \int_e \{K\bar{\sigma}_i^h \cdot \mathbf{n}_e\} [\bar{u}_i^h], \\ II &= 2k \sum_{i=1}^m (1 - \kappa) \sum_{e \in \Gamma_h \cup \Gamma_D} \int_e \{D\nabla \bar{u}_i^h \cdot \mathbf{n}_e\} [\bar{u}_i^h]. \end{aligned}$$

Now, using  $\|\bar{\sigma}_i^h \cdot \mathbf{n}_e\|_{0, \partial E} \leq Ch^{-1/2} \|\bar{\sigma}_i^h\|_{0, E}$  and recalling that  $|e| \leq Ch^{d-1}$ , for  $I$  we have,

$$\begin{aligned} |I| &\leq 2k \sum_{i=1}^m \sum_{e \in \Gamma_h \cup \Gamma_D} K \|\{\bar{\sigma}_i^h \cdot \mathbf{n}_e\}\|_{0, e} \|\bar{u}_i^h\|_{0, e}, \\ &\leq 2\epsilon_1 k \sum_{i=1}^m \sum_{e \in \Gamma_h \cup \Gamma_D} K^2 \left( \frac{|e|^\beta}{\delta} \right) \|\{\bar{\sigma}_i^h \cdot \mathbf{n}_e\}\|_{0, e}^2 + \frac{k}{2\epsilon_1} \sum_{i=1}^m \sum_{e \in \Gamma_h \cup \Gamma_D} \left( \frac{\delta}{|e|^\beta} \right) \|\bar{u}_i^h\|_{0, e}^2, \\ &\leq 2\epsilon_1 k \sum_{i=1}^m \frac{K^2 Ch^{(d-1)\beta-1}}{\delta} \|\bar{\sigma}_i^h\|_0^2 + \frac{k}{2\epsilon_1} \sum_{i=1}^m J_0^{\delta, \beta}(\bar{u}_i^h, \bar{u}_i^h). \end{aligned}$$

Similarly, since  $|\kappa - 1| \leq 2$ ,

$$\begin{aligned} |II| &\leq 4k \sum_{i=1}^m \sum_{e \in \Gamma_h \cup \Gamma_D} \left( \frac{|e|^\beta}{\delta} \right)^{1/2} \|\{D\nabla \bar{u}_i^h \cdot \mathbf{n}_e\}\|_{0, e} \left( \frac{\delta}{|e|^\beta} \right)^{1/2} \|\bar{u}_i^h\|_{0, e}, \\ &\leq 2\epsilon_2 k \sum_{i=1}^m \sum_E \frac{DC h^{(d-1)\beta-1}}{\delta} \|D^{1/2} \nabla \bar{u}_i^h\|_{0, E}^2 + \frac{2k}{\epsilon_2} \sum_{i=1}^m J_0^{\delta, \beta}(\bar{u}_i^h, \bar{u}_i^h). \end{aligned}$$

With these we arrive at,

$$\begin{aligned} & \|u_m^h\|_0^2 + \frac{K}{\mu} \|\sigma_m^h\|_0^2 + \left(2 - \frac{1}{2\epsilon_1} - \frac{2}{\epsilon_2}\right) k \sum_{i=1}^m \|\bar{u}_i^h\|_{\mathcal{A}}^2 + 2k \sum_{i=1}^m \frac{K}{\mu} (\gamma(\mathcal{B}_{i,n} u^h) \bar{\sigma}_i^h, \bar{\sigma}_i^h) \\ & \leq \|u_0^h\|_0^2 + \frac{K}{\mu} \|\sigma_0^h\|_0^2 + 2k \left| \sum_{i=1}^m L_i(\bar{u}_i^h) \right| + 2k \sum_{i=1}^m \frac{C h^{(d-1)\beta-1}}{\delta} \left( K^2 \epsilon_1 \|\bar{\sigma}_i^h\|_0^2 + D \epsilon_2 \|\bar{u}_i^h\|_{\mathcal{A}}^2 \right). \end{aligned}$$

Now, using Lemma 2.1,

$$\begin{aligned} 2k \left| \sum_{i=1}^m L_i(\bar{u}_i^h) \right| &= k \left| \sum_{i=1}^m \left( L(t_i; \bar{u}_i^h) + L(t_{i-1}; \bar{u}_i^h) \right) \right|, \\ &= k \left| \sum_{i=1}^m \left( (f(t_i), \bar{u}_i^h) + (f(t_{i-1}), \bar{u}_i^h) + (g(t_i), \bar{u}_i^h)_{\Gamma_N} + (g(t_{i-1}), \bar{u}_i^h)_{\Gamma_N} \right) \right|, \\ &= 2k \left| \sum_{i=1}^m \left( (\bar{f}_i, \bar{u}_i^h) + (\bar{g}_i, \bar{u}_i^h)_{\Gamma_N} \right) \right|, \\ &\leq 2k \sum_{i=1}^m C_f \|\bar{f}_i\|_0 \|\bar{u}_i^h\|_{\mathcal{A}} + 2k \sum_{i=1}^m C_g h^{-1/2} \|\bar{g}_i\|_{0, \Gamma_N} \|\bar{u}_i^h\|_{\mathcal{A}}, \\ &\leq \epsilon_3 k \sum_{i=1}^m C_f^2 \|\bar{f}_i\|_0^2 + \epsilon_3 k \sum_{i=1}^m C_g^2 h^{-1} \|\bar{g}_i\|_{0, \Gamma_N}^2 + \frac{2k}{\epsilon_3} \sum_{i=1}^m \|\bar{u}_i^h\|_{\mathcal{A}}^2. \end{aligned}$$

With this and (6), we now have,

$$\begin{aligned} & \|u_m^h\|_0^2 + \frac{K}{\mu} \|\sigma_m^h\|_0^2 + \left(2 - \frac{1}{2\epsilon_1} - \frac{2}{\epsilon_2} - \frac{2}{\epsilon_3}\right) k \sum_{i=1}^m \|\bar{u}_i^h\|_{\mathcal{A}}^2 + 2k \sum_{i=1}^m \frac{K \gamma_G}{\mu} \|\bar{\sigma}_i^h\|_0^2 \\ & \leq \|u_0^h\|_0^2 + \frac{K}{\mu} \|\sigma_0^h\|_0^2 + \epsilon_3 k \sum_{i=1}^m \left( C_f^2 \|\bar{f}_i\|_0^2 + C_g^2 h^{-1} \|\bar{g}_i\|_{0, \Gamma_N}^2 \right) \\ & \quad + 2k \sum_{i=1}^m \frac{C h^{(d-1)\beta-1}}{\delta} \left( K^2 \epsilon_1 \|\bar{\sigma}_i^h\|_0^2 + D \epsilon_2 \|\bar{u}_i^h\|_{\mathcal{A}}^2 \right). \end{aligned}$$

Setting  $\epsilon_2 = \epsilon_3 = 6$  and  $\epsilon_1 = 3/2$  means that we can write this as,

$$\begin{aligned} & \|u_m^h\|_0^2 + \frac{K}{\mu} \|\sigma_m^h\|_0^2 + \left(1 - \frac{12DC\hat{h}^{(d-1)\beta-1}}{\delta}\right) k \sum_{i=1}^m \|\bar{u}_i^h\|_{\mathcal{A}}^2 \\ & \quad + \left(\frac{\gamma_G}{\mu} - \frac{3KC\hat{h}^{(d-1)\beta-1}}{2\delta}\right) 2Kk \sum_{i=1}^m \|\bar{\sigma}_i^h\|_0^2 \\ & \leq \|u_0^h\|_0^2 + \frac{K}{\mu} \|\sigma_0^h\|_0^2 + 6k \sum_{i=1}^m \left( C_f^2 \|\bar{f}_i\|_0^2 + C_g^2 h^{-1} \|\bar{g}_i\|_{0, \Gamma_N}^2 \right), \end{aligned}$$

and choosing some positive constant  $C^* < \min\{1, \gamma_G/\mu\}$ , and requiring that

$$\delta \geq 3C\hat{h}^{(d-1)\beta-1} \max \left\{ \frac{4D}{1 - C^*}, \frac{\mu K}{2\gamma_G - 2\mu C^*} \right\},$$

we arrive at the theorem.  $\square$ .

Since this is a finite dimensional problem, we can infer existence from uniqueness in the linear case where  $n = 2$ . Since this is the more practical of the two algorithms we are content with this. Also, at least for the original model of Cohen *et al.*, [6], it seems from [2] that such analysis for the nonlinear problem is highly non-trivial.

**Theorem 2.3 (discrete existence and uniqueness)** *Under the conditions of Theorem 2.2, the discrete solution exists for  $n = 2$  and is unique.*

**Remark 2.4** *The condition that  $\delta$  'be large enough' in Theorem 2.2 can be removed in the non-symmetric case,  $\kappa = 1$ , by requiring a small enough time step,  $k$ . To see this note that the term  $II$  in the proof vanishes and that the second term in the bound for  $I$  can be moved to the left with an appropriate choice of  $\epsilon_1$ . After applying the triangle inequality to  $\|\bar{\sigma}_i^h\|_0^2$ , the term  $\|\sigma_m^h\|_0^2$  can also be moved to the left if  $k$  is small enough, and the remaining terms are bounded by a discrete Gronwall inequality.*

### 3 Error estimate

In this section we derive error estimates for our schemes encompassing the cases  $\kappa = \pm 1$  and  $n = 1$  or  $2$ . First we need some standard Taylor's series estimates, and it is convenient to define,

$$\Delta_i v := \frac{v_t(t_i) + v_t(t_{i-1})}{2} - \frac{v(t_i) - v(t_{i-1})}{k},$$

which we recognise as (the negative of) the error in the trapezium rule.

**Lemma 3.1 (Taylor estimates)** *Whenever  $v$  has the indicated regularity we have positive constants,  $C$ , independent of  $h$  and  $k$  such that,*

$$\|v(t_{i-1/2}) - \bar{v}_i\|_0 \leq Ck^{3/2} \|v_{tt}\|_{L_2(t_{i-1}, t_i; L_2(\Omega))}, \quad (23)$$

$$\|v(k/2) - v(0)\|_0 \leq Ck \|v_t\|_{L_\infty(0, k/2; L_2(\Omega))}, \quad (24)$$

$$\left\| v(t_{i-1/2}) - \frac{3v(t_{i-1}) - v(t_{i-2})}{2} \right\|_0 \leq Ck^{3/2} \|v_{tt}\|_{L_2(t_{i-2}, t_{i-1/2}; L_2(\Omega))}, \quad (25)$$

$$\|v_t(t_{i-1/2}) - \partial_t v_i\|_0 \leq Ck^{3/2} \|v_{ttt}\|_{L_2(t_{i-1}, t_i; L_2(\Omega))}, \quad (26)$$

and,

$$\|\Delta_i v\|_0 \leq Ck^{3/2} \|v_{ttt}\|_{L_2(t_{i-1}, t_i; L_2(\Omega))}, \quad (27)$$

from the Peano kernel theorem applied to the trapezoidal rule for numerical integration.

We define,

$$\begin{aligned} \chi_i &:= u_i^h - u^\perp(t_i), & \eta_i &:= \sigma_i^h - \sigma^*(t_i), \\ \xi(t_i) &:= u(t_i) - u^\perp(t_i), & \theta(t_i) &:= \sigma(t_i) - \sigma^*(t_i), \end{aligned}$$

where  $\sigma^* \in \mathcal{D}_{r-1}(\mathcal{E}_h)$  is the nodal interpolant to  $\sigma$ , and  $u^\perp \in \mathcal{D}_r(\mathcal{E}_h)$  is the elliptic projection of  $u$  defined by,

$$A(u^\perp, v) = A(u, v) \quad \forall v \in \mathcal{D}_r(\mathcal{E}_h). \quad (28)$$

**Proposition 3.2 (estimates for the elliptic projection)** *If  $u \in C(\bar{\Omega})$  and  $u^\perp \in \mathcal{D}_r(\mathcal{E}_h)$  is defined through (28) for  $\kappa = \pm 1$ , we have for  $m = 0, 1, 2, \dots$  and  $t \geq 0$  that,*

$$\left\| \frac{\partial^m}{\partial t^m} (u(t) - u^\perp(t)) \right\|_{\mathcal{A}} \leq Ch^s \left\| \frac{\partial^m u}{\partial t^m}(t) \right\|_{s+1}, \quad (29)$$

$$\left\| \frac{\partial^m}{\partial t^m} (u(t) - u^\perp(t)) \right\|_0 \leq Ch^s \left\| \frac{\partial^m u}{\partial t^m}(t) \right\|_{s+1}, \quad (30)$$

$$\left\| \frac{\partial^m u^\perp}{\partial t^m}(t) \right\|_{\mathcal{A}} \leq C \left\| \frac{\partial^m u}{\partial t^m}(t) \right\|_2, \quad (31)$$

whenever  $\partial^m u(t)/\partial t^m \in H^{s+1}(\Omega)$  and  $1 \leq s \leq r$ .

When  $m = 0$  the proof of (29) is given in [15] (the ‘NIPG’ scheme) for the non-symmetric case,  $\kappa = 1$ , and can be readily established for  $\kappa = -1$  by similar arguments. The non-optimal (30) then follows from (29) and (18) (an optimal  $L_2$  estimate is also given in [15], but we don’t need it here). The stability estimate, (31), follows from,

$$\|u^\perp(t)\|_{\mathcal{A}} \leq \|u(t) - u^\perp(t)\|_{\mathcal{A}} + \|u(t)\|_{\mathcal{A}},$$

along with (29) (with  $s = 1$ ) and the fact that  $[u(t)] = 0$ . The estimates then follow for  $m \geq 1$  by differentiating (28).

For use later, we note also that,

$$\|\boldsymbol{\sigma}^*(t)\|_{\mathbf{L}_\infty(\Omega)} \leq C\|\boldsymbol{\sigma}(t)\|_{\mathbf{L}_\infty(\Omega)}. \quad (32)$$

The next result is a lemma that deals with the error generated by the nonlinear term.

**Lemma 3.3 (nonlinearity error)** *For  $n = 1$  or  $2$  we have,*

$$\begin{aligned} & \left| \left( \overline{(\gamma(u)\boldsymbol{\sigma})_i} \right) - \gamma(\mathcal{B}_{i,n}u^h)\bar{\boldsymbol{\sigma}}_i^*, \bar{\boldsymbol{\eta}}_i \right| \\ & \leq \frac{Ch^{2r}}{\epsilon} \left( \|\boldsymbol{\sigma}\|_{L_\infty(0,T;\mathbf{H}^r(\Omega))}^2 + \|\boldsymbol{\sigma}\|_{L_\infty(0,T;\mathbf{L}_\infty(\Omega))}^2 \|u\|_{L_\infty(0,T;\mathbf{H}^{r+1}(\Omega))}^2 \right) \\ & \quad + \frac{Ck^3}{\epsilon} \left( \|(\gamma(u)\boldsymbol{\sigma})_{tt}\|_{L_2(t_{i-1},t_i;\mathbf{L}_2(\Omega))}^2 + \|\boldsymbol{\sigma}_{tt}\|_{L_2(t_{i-1},t_i;\mathbf{L}_2(\Omega))}^2 \right) \\ & \quad + \frac{C}{\epsilon} \|\boldsymbol{\sigma}\|_{L_\infty(0,T;\mathbf{L}_\infty(\Omega))}^2 \|\chi_{i-1}\|_0^2 \\ & + \begin{cases} \frac{C}{\epsilon} \|\boldsymbol{\sigma}\|_{L_\infty(0,T;\mathbf{L}_\infty(\Omega))}^2 \left( k^3 \|u_{tt}\|_{L_2(t_{i-1},t_i;\mathbf{H}^2(\Omega))}^2 + \|\chi_i\|_0^2 \right) + \frac{\gamma G \epsilon}{2} \|\bar{\boldsymbol{\eta}}_i\|_0^2 \\ \text{for } n = 1, i = 1, \dots, N, \\ \\ \frac{Ck^2}{\epsilon} \|\boldsymbol{\sigma}\|_{L_\infty(0,T;\mathbf{L}_\infty(\Omega))}^2 \|u_t\|_{L_\infty(0,k/2;\mathbf{H}^2(\Omega))}^2 + \frac{\gamma G \epsilon}{2} \|\boldsymbol{\eta}_1\|_0^2 + \frac{\gamma G \epsilon}{2} \|\boldsymbol{\eta}_0\|_0^2 \\ \text{for } n = 2, i = 1, \\ \\ \frac{C}{\epsilon} \|\boldsymbol{\sigma}\|_{L_\infty(0,T;\mathbf{L}_\infty(\Omega))}^2 \left( k^3 \|u_{tt}\|_{L_2(t_{i-2},t_{i-1/2};\mathbf{H}^2(\Omega))}^2 + \|\chi_{i-2}\|_0^2 \right) + \frac{\gamma G \epsilon}{2} \|\bar{\boldsymbol{\eta}}_i\|_0^2 \\ \text{for } n = 2, i = 2, \dots, N, \end{cases} \end{aligned}$$

for a constant  $C$  independent of  $h, k$  and  $\epsilon$  and for all  $\epsilon > 0$ .

**Proof.** We have, from (23) in Lemma 3.1,

$$\begin{aligned} & \left| \left( \overline{(\gamma(u)\boldsymbol{\sigma}(t_i))} \right) - \gamma(\mathcal{B}_{i,n}u^h)\bar{\boldsymbol{\sigma}}_i^*, \bar{\boldsymbol{\eta}}_i \right| \leq \|\bar{\boldsymbol{\eta}}_i\|_0 \left( \|\overline{(\gamma(u)\boldsymbol{\sigma}(t_i))} - \gamma(u(t_{i-1/2}))\boldsymbol{\sigma}(t_{i-1/2})\|_0 \right. \\ & \quad \left. + \|\gamma(u(t_{i-1/2}))\boldsymbol{\sigma}(t_{i-1/2}) - \gamma(\mathcal{B}_{i,n}u^h)\bar{\boldsymbol{\sigma}}_i^*\|_0 \right), \\ & \leq Ck^{3/2} \|(\gamma(u)\boldsymbol{\sigma})_{tt}\|_{L_2(t_{i-1},t_i;\mathbf{L}_2(\Omega))} \|\bar{\boldsymbol{\eta}}_i\|_0 \\ & \quad + \|\bar{\boldsymbol{\eta}}_i\|_0 \left( \|\gamma(u(t_{i-1/2}))(\boldsymbol{\sigma}(t_{i-1/2}) - \bar{\boldsymbol{\sigma}}_i^*)\|_0 \right. \\ & \quad \left. + \|\gamma(u(t_{i-1/2})) - \gamma(\mathcal{B}_{i,n}u^h)\|_0 \|\bar{\boldsymbol{\sigma}}_i^*\|_{\mathbf{L}_\infty(\Omega)} \right), \\ & \leq Ck^{3/2} \|(\gamma(u)\boldsymbol{\sigma})_{tt}\|_{L_2(t_{i-1},t_i;\mathbf{L}_2(\Omega))} \|\bar{\boldsymbol{\eta}}_i\|_0 \\ & \quad + \gamma_R \|\bar{\boldsymbol{\eta}}_i\|_0 \left( \|\boldsymbol{\sigma}(t_{i-1/2}) - \bar{\boldsymbol{\sigma}}_i\|_0 + \|\bar{\boldsymbol{\sigma}}_i - \bar{\boldsymbol{\sigma}}_i^*\|_0 \right) \\ & \quad + C'_\gamma \|\bar{\boldsymbol{\eta}}_i\|_0 \|\bar{\boldsymbol{\sigma}}_i^*\|_{\mathbf{L}_\infty(\Omega)} \|u(t_{i-1/2}) - \mathcal{B}_{i,n}u^h\|_0, \end{aligned}$$

where we observed, using (8), that,

$$\begin{aligned} \|\gamma(u(t_{i-1/2})) - \gamma(\mathcal{B}_{i,n}u^h)\|_0 &= \left\| \int_0^1 \gamma'(su(t_{i-1/2}) + (1-s)\mathcal{B}_{i,n}u^h) ds (u(t_{i-1/2}) - \mathcal{B}_{i,n}u^h) \right\|_0 \\ &\leq C'_\gamma \|u(t_{i-1/2}) - \mathcal{B}_{i,n}u^h\|_0. \end{aligned}$$

Using (23), (24) and (25) from Lemma 3.1, along with (16) and (32) we therefore arrive at,

$$\begin{aligned} \left| \left( \overline{(\gamma(u)\boldsymbol{\sigma}(t_i))} - \gamma(\mathcal{B}_{i,n}u^h)\bar{\boldsymbol{\sigma}}_i^*, \bar{\boldsymbol{\eta}}_i \right) \right| &\leq Ck^{3/2} \|(\gamma(u)\boldsymbol{\sigma})_{tt}\|_{L_2(t_{i-1}, t_i; \mathbf{L}_2(\Omega))} \|\bar{\boldsymbol{\eta}}_i\|_0 \\ &\quad + \gamma_R \|\bar{\boldsymbol{\eta}}_i\|_0 \left( Ck^{3/2} \|\boldsymbol{\sigma}_{tt}\|_{L_2(t_{i-1}, t_i; \mathbf{L}_2(\Omega))} + Ch^r \|\boldsymbol{\sigma}\|_{L_\infty(0, T; \mathbf{H}^r(\Omega))} \right) \\ &\quad + C \|\bar{\boldsymbol{\eta}}_i\|_0 \|\boldsymbol{\sigma}\|_{L_\infty(0, T; \mathbf{L}_\infty(\Omega))} \|u(t_{i-1/2}) - \mathcal{B}_{i,n}u^h\|_0. \end{aligned}$$

Now, using Proposition 3.2,

$$\begin{aligned} \|u(t_{i-1/2}) - \mathcal{B}_{i,n}u^h\|_0 &\leq \|u(t_{i-1/2}) - u^\perp(t_{i-1/2})\|_0 + \|u^\perp(t_{i-1/2}) - \mathcal{B}_{i,n}u^\perp\|_0 \\ &\quad + \|\mathcal{B}_{i,n}u^\perp - \mathcal{B}_{i,n}u^h\|_0, \\ &\leq Ch^r \|u\|_{L_\infty(0, T; H^{r+1}(\Omega))} + \|u^\perp(t_{i-1/2}) - \mathcal{B}_{i,n}u^\perp\|_0 + \|\mathcal{B}_{i,n}\chi\|_0, \end{aligned}$$

and this is as far as we can get without distinguishing between  $n = 1$  and  $n = 2$ .

So, firstly, for  $n = 1$  we have,

$$\begin{aligned} \|u(t_{i-1/2}) - \mathcal{B}_{i,1}u^h\|_0 &\leq Ch^r \|u\|_{L_\infty(0, T; H^{r+1}(\Omega))} + Ck^{3/2} \|u_{tt}\|_{L_2(t_{i-1}, t_i; H^2(\Omega))} \\ &\quad + \frac{1}{2} \|\chi_i\|_0 + \frac{1}{2} \|\chi_{i-1}\|_0, \end{aligned}$$

where we used (23) from Lemma 3.1 and (30) with  $m = 0$ .

Secondly, using (24) from Lemma 3.1, in the case  $n = 2$  we have when  $i = 1$  that,

$$\|u(t_{i-1/2}) - \mathcal{B}_{i,2}u^h\|_0 \leq Ch^r \|u\|_{L_\infty(0, T; H^{r+1}(\Omega))} + Ck \|u_t\|_{L_\infty(0, k/2; H^2(\Omega))} + \|\chi_0\|_0,$$

while if  $i > 1$ , with (25) from Lemma 3.1,

$$\begin{aligned} \|u(t_{i-1/2}) - \mathcal{B}_{i,2}u^h\|_0 &\leq Ch^r \|u\|_{L_\infty(0, T; H^{r+1}(\Omega))} + Ck^{3/2} \|u_{tt}\|_{L_2(t_{i-2}, t_{i-1/2}; H^2(\Omega))} \\ &\quad + \frac{3}{2} \|\chi_{i-1}\|_0 + \frac{1}{2} \|\chi_{i-2}\|_0. \end{aligned}$$

Assembling these estimates then gives,

$$\begin{aligned} &\left| \left( \overline{(\gamma(u)\boldsymbol{\sigma})_i} - \gamma(\mathcal{B}_{i,n}u^h)\bar{\boldsymbol{\sigma}}_i^*, \bar{\boldsymbol{\eta}}_i \right) \right| \\ &\leq Ck^{3/2} \left( \|(\gamma(u)\boldsymbol{\sigma})_{tt}\|_{L_2(t_{i-1}, t_i; \mathbf{L}_2(\Omega))} + \gamma_R \|\boldsymbol{\sigma}_{tt}\|_{L_2(t_{i-1}, t_i; \mathbf{L}_2(\Omega))} \right) \|\bar{\boldsymbol{\eta}}_i\|_0 \\ &\quad + Ch^r \left( \|\boldsymbol{\sigma}\|_{L_\infty(0, T; \mathbf{L}_\infty(\Omega))} \|u\|_{L_\infty(0, T; H^{r+1}(\Omega))} + \gamma_R \|\boldsymbol{\sigma}\|_{L_\infty(0, T; \mathbf{H}^r(\Omega))} \right) \|\bar{\boldsymbol{\eta}}_i\|_0 \\ &\quad + \|\bar{\boldsymbol{\eta}}_i\|_0 \|\boldsymbol{\sigma}\|_{L_\infty(0, T; \mathbf{L}_\infty(\Omega))} \times \begin{cases} Ck^{3/2} \|u_{tt}\|_{L_2(t_{i-1}, t_i; H^2(\Omega))} + \frac{1}{2} \|\chi_i\|_0 + \frac{1}{2} \|\chi_{i-1}\|_0, \\ \quad \text{for } n = 1, i \geq 1; \\ \\ Ck \|u_t\|_{L_\infty(0, k/2; H^2(\Omega))} + \|\chi_0\|_0, \\ \quad \text{for } n = 2, i = 1; \\ \\ Ck^{3/2} \|u_{tt}\|_{L_2(t_{i-2}, t_{i-1/2}; H^2(\Omega))} \\ \quad + \frac{3}{2} \|\chi_{i-1}\|_0 + \frac{1}{2} \|\chi_{i-2}\|_0, \\ \quad \text{for } n = 2, i > 1. \end{cases} \end{aligned}$$

Several applications of Young's inequality then completes the proof.  $\square$

Before giving the error estimate we recall, from e.g. [1, Theorem 4.12], that if  $\Omega \subset \mathbb{R}^d$ , for  $d = 2$  or  $3$ , satisfies a cone condition then  $\|v\|_{L_\infty(\Omega)} \leq C\|v\|_m$  for  $m > d/2$ . Moreover,

$$H^1(\Omega) \hookrightarrow L_q(\Omega) \quad \text{for} \quad \begin{cases} 2 \leq q < \infty & \text{if } d = 2; \\ 2 \leq q \leq 6 & \text{if } d = 3, \end{cases} \quad (33)$$

and then  $\|v\|_{L_q(\Omega)} \leq C\|v\|_1$  for all  $v \in H^1(\Omega)$ . Also, if  $(X, \|\cdot\|_X)$  is a Banach space then, for  $v: I \rightarrow X$ , we have,

$$\|v\|_{L_\infty(0,\tau;X)} \leq C(\tau) \left( \|v(0)\|_X + \|v_t\|_{L_p(0,\tau;X)} \right) \quad \forall \tau \in \bar{I} \quad (34)$$

and for  $1 \leq p \leq \infty$ .

Now we can state the error estimate. The regularity requirements stated in this are given simply as they appear in the proof and in Lemma 3.3. We return to this point later.

**Theorem 3.4 (error estimate)** *Let  $\hat{h} \leq \text{diam}(\Omega)$  and  $\hat{k} \leq T$  be positive constants and for  $r \geq 1$  assume that,  $\check{u} \in H^{r+1}(\Omega)$ ,  $\check{\sigma} \in \mathbf{H}^r(\Omega)$ ,*

- $u \in W_\infty^1(I; H^{r+1}(\Omega)) \cap H^2(I; H^2(\Omega)) \cap H^3(I; L_2(\Omega))$ ,
- $\sigma \in L_\infty(I; \mathbf{L}_\infty(\Omega)) \cap W_\infty^1(I; \mathbf{H}^r(\Omega)) \cap H^3(I; \mathbf{L}_2(\Omega))$ ,
- $(\gamma(u)\sigma)_{tt} \in L_2(I; \mathbf{L}_2(\Omega))$ ,

then for  $\beta \geq (d-1)^{-1}$ ,  $h \leq \hat{h}$ ,  $\hat{h}^{(d-1)\beta-1}/\delta$  small enough (for  $n = 1$  and  $2$ ) and  $k \leq \hat{k}$ , where  $\hat{k}$  is small enough (for  $n = 1$  only), we have a positive constant,  $C$ , independent of  $h$  and  $k$  such that,

$$\|u(t_m) - u_m^h\|_0 + \|\sigma(t_m) - \sigma_m^h\|_0 + \left( k \sum_{i=1}^m \|\bar{u}_i - \bar{u}_i^h\|_{\mathcal{A}}^2 + \|\bar{\sigma}_i - \bar{\sigma}_i^h\|_0^2 \right)^{1/2} \leq C(h^r + k^2)$$

for each  $m = 1, \dots, N$ .

**Proof.** We average (19) between  $t_i$  and  $t_{i-1}$  and subtract it from (21), and do the same with (20) and (22). Adding the results of these then gives an error equation,

$$\begin{aligned} & (\partial_t \chi_i, v) + (\partial_t \eta_i, \mathbf{w}) + A(\bar{\chi}_i, v) = (\Delta_i u, v) + (\Delta_i \sigma, \mathbf{w}) + (\partial_t \xi_i, v) + (\partial_t \theta_i, \mathbf{w}) + A(\bar{\xi}_i, v) \\ & - \sum_E (K \bar{\eta}_i, \nabla v)_E + \sum_E (K \bar{\theta}_i, \nabla v)_E \\ & + \sum_E (\mu \nabla \bar{\chi}_i, \mathbf{w})_E - \sum_E (\mu \nabla \bar{\xi}_i, \mathbf{w})_E \\ & + \sum_{e \in \Gamma_h \cup \Gamma_D} \int_e \{K \bar{\eta}_i \cdot \mathbf{n}_e\} [v] - \sum_{e \in \Gamma_h \cup \Gamma_D} \int_e \{K \bar{\theta}_i \cdot \mathbf{n}_e\} [v] \\ & + \overline{(\gamma(u)\sigma(t_i))} - \gamma(\mathcal{B}_{i,n} u^h) \bar{\sigma}_i^h, \mathbf{w} \quad \forall v \in \mathcal{D}_r(\mathcal{E}_h) \text{ and } \forall \mathbf{w} \in \mathcal{D}_{r-1}(\mathcal{E}_h). \end{aligned}$$

We now choose  $v = \bar{\chi}_i$  and  $\mathbf{w} = (K/\mu)\bar{\boldsymbol{\eta}}_i$ , multiply by  $2k$  and sum over  $i = 1, \dots, m \leq N$  to get,

$$\begin{aligned}
& \|\chi_m\|_0^2 + \frac{K}{\mu} \|\boldsymbol{\eta}_m\|_0^2 + 2k \sum_{i=1}^m \|\bar{\chi}_i\|_{\mathcal{A}}^2 + 2k \sum_{i=1}^m \frac{K}{\mu} (\gamma(\mathcal{B}_{i,n} u^h) \bar{\boldsymbol{\eta}}_i, \bar{\boldsymbol{\eta}}_i) \\
&= \|\chi_0\|_0^2 + \frac{K}{\mu} \|\boldsymbol{\eta}_0\|_0^2 + 2k \sum_{i=1}^m (\Delta_i u, \bar{\chi}_i) + 2k \sum_{i=1}^m \frac{K}{\mu} (\Delta_i \boldsymbol{\sigma}, \bar{\boldsymbol{\eta}}_i) \\
&\quad + 2k \sum_{i=1}^m (\partial_t \xi_i, \bar{\chi}_i) + 2k \sum_{i=1}^m A(\bar{\xi}_i, \bar{\chi}_i) + \frac{2Kk}{\mu} \sum_{i=1}^m (\partial_t \boldsymbol{\theta}_i, \bar{\boldsymbol{\eta}}_i) \\
&\quad + 2k \sum_{i=1}^m \sum_E (K \bar{\boldsymbol{\theta}}_i, \nabla \bar{\chi}_i)_E - 2k \sum_{i=1}^m \sum_E (K \nabla \bar{\xi}_i, \bar{\boldsymbol{\eta}}_i)_E \\
&\quad + 2k \sum_{i=1}^m (1 - \kappa) \sum_{e \in \Gamma_h \cup \Gamma_D} \int_e \{D \nabla \bar{\chi}_i \cdot \mathbf{n}_e\} [\bar{\chi}_i] \\
&\quad + 2k \sum_{i=1}^m \sum_{e \in \Gamma_h \cup \Gamma_D} \int_e \{K \bar{\boldsymbol{\eta}}_i \cdot \mathbf{n}_e\} [\bar{\chi}_i] - 2k \sum_{i=1}^m \sum_{e \in \Gamma_h \cup \Gamma_D} \int_e \{K \bar{\boldsymbol{\theta}}_i \cdot \mathbf{n}_e\} [\bar{\chi}_i] \\
&\quad + \frac{2Kk}{\mu} \sum_{i=1}^m (\gamma(u) \boldsymbol{\sigma}(t_i) - \gamma(\mathcal{B}_{i,n} u^h) \boldsymbol{\sigma}_i^*, \bar{\boldsymbol{\eta}}_i), \\
&= T_1 + \dots + T_{13}.
\end{aligned}$$

We now take each term in turn. By the  $L_2(\Omega)$  projection we have,  $(\chi_0, v) = (\xi(0), v)$  for all  $v \in \mathcal{D}_r(\mathcal{E}_h)$ , which, from (30), results in,

$$|T_1| = \|\chi_0\|_0^2 \leq \|\xi(0)\|_0^2 \leq Ch^{2r} \|\check{u}\|_{r+1}^2.$$

Similarly, we have  $(\boldsymbol{\eta}_0, \mathbf{w}) = (\boldsymbol{\theta}(0), \mathbf{w})$  for all  $\mathbf{w} \in \mathcal{D}_{r-1}(\mathcal{E}_h)$  and, from (16), this gives,

$$|T_2| = \|\boldsymbol{\eta}_0\|_0^2 \leq \|\boldsymbol{\theta}(0)\|_0^2 \leq Ch^{2r} \|\check{\boldsymbol{\sigma}}\|_r^2.$$

For  $T_3$  and  $T_4$  we appeal to (27) from Lemma 3.1 and (18) to get,

$$|T_3| \leq \frac{Ck}{\epsilon_3} \sum_{i=1}^m \|\Delta_i u\|_0^2 + \epsilon_3 k \sum_{i=1}^m \|\bar{\chi}_i\|_{\mathcal{A}}^2 \leq \frac{Ck^4}{\epsilon_3} \|u_{ttt}\|_{L_2(0, t_m; L_2(\Omega))}^2 + \epsilon_3 k \sum_{i=1}^m \|\bar{\chi}_i\|_{\mathcal{A}}^2,$$

and,

$$\begin{aligned}
|T_4| &\leq \frac{Kk}{\mu \gamma_G \epsilon_4} \sum_{i=1}^m \|\Delta_i \boldsymbol{\sigma}\|_0^2 + \epsilon_4 k \sum_{i=1}^m \frac{K \gamma_G}{\mu} \|\bar{\boldsymbol{\eta}}_i\|_0^2, \\
&\leq \frac{Ck^4}{\epsilon_4} \|\boldsymbol{\sigma}_{ttt}\|_{L_2(0, t_m; L_2(\Omega))}^2 + \epsilon_4 k \sum_{i=1}^m \frac{K \gamma_G}{\mu} \|\bar{\boldsymbol{\eta}}_i\|_0^2.
\end{aligned}$$

Using (18), (30) and (26) from Lemma 3.1, we have for  $T_5$  that,

$$\begin{aligned}
|T_5| &\leq \frac{Ck}{\epsilon_5} \sum_{i=1}^m (\|\partial_t \xi_i - \xi_t(t_{i-1/2})\|_0^2 + \|\xi_t(t_{i-1/2})\|_0^2) + \epsilon_5 k \sum_{i=1}^m \|\bar{\chi}_i\|_{\mathcal{A}}^2, \\
&\leq \frac{Ck^4}{\epsilon_5} \|u_{ttt}\|_{L_2(0, t_m; H^2(\Omega))}^2 + \frac{Ct_m h^{2r}}{\epsilon_5} \|u_t\|_{L_\infty(0, t_m; H^{r+1}(\Omega))}^2 + \epsilon_5 k \sum_{i=1}^m \|\bar{\chi}_i\|_{\mathcal{A}}^2,
\end{aligned}$$

where we used (18) and (31) to get  $\|\xi_{ttt}\|_0 \leq C \|u_{ttt}\|_2$ .

Now,  $T_6 = 0$  from (28) and for  $T_7$  we argue similarly as for  $T_5$  and obtain,

$$\begin{aligned} |T_7| &\leq \frac{Kk}{\mu\gamma_G\epsilon_7} \sum_{i=1}^m \|\partial_t \boldsymbol{\theta}_i\|_0^2 + \epsilon_7 k \sum_{i=1}^m \frac{K\gamma_G}{\mu} \|\bar{\boldsymbol{\eta}}_i\|_0^2, \\ &\leq \frac{Ck^4}{\epsilon_7} \|\boldsymbol{\sigma}_{ttt}\|_{L_2(0,t_m;L_2(\Omega))} + \frac{Ct_m h^{2r}}{\epsilon_7} \|\boldsymbol{\sigma}_t\|_{L_\infty(0,t_m;H^r(\Omega))}^2 + \epsilon_7 k \sum_{i=1}^m \frac{K\gamma_G}{\mu} \|\bar{\boldsymbol{\eta}}_i\|_0^2, \end{aligned}$$

where we used the estimate  $\|\boldsymbol{\theta}_{ttt}\|_0 \leq C\|\boldsymbol{\sigma}_{ttt}\|_0$ . For  $T_8$ ,

$$|T_8| \leq \frac{Ck}{\epsilon_8 D^2} \sum_{i=1}^m \|\bar{\boldsymbol{\theta}}_i\|_0^2 + \epsilon_8 k \sum_{i=1}^m \|\bar{\chi}_i\|_{\mathcal{A}}^2 \leq \frac{Ct_m h^{2r}}{\epsilon_8} \|\boldsymbol{\sigma}\|_{L_\infty(0,t_m;H^r(\Omega))}^2 + \epsilon_8 k \sum_{i=1}^m \|\bar{\chi}_i\|_{\mathcal{A}}^2,$$

and  $T_9$ ,

$$\begin{aligned} |T_9| &\leq 2k \sum_{i=1}^m \frac{K}{D} \|\bar{\xi}_i\|_{\mathcal{A}} \|\bar{\boldsymbol{\eta}}_i\|_0 \leq \frac{k}{\epsilon_9} \sum_{i=1}^m \frac{\mu K}{\gamma_G D^2} \|\bar{\xi}_i\|_{\mathcal{A}}^2 + \epsilon_9 k \sum_{i=1}^m \frac{K\gamma_G}{\mu} \|\bar{\boldsymbol{\eta}}_i\|_0^2, \\ &\leq \frac{Ct_m h^{2r}}{\epsilon_9} \|u\|_{L_\infty(0,t_m;H^{r+1}(\Omega))}^2 + \epsilon_9 k \sum_{i=1}^m \frac{K\gamma_G}{\mu} \|\bar{\boldsymbol{\eta}}_i\|_0^2. \end{aligned}$$

We now note that  $T_{10} = 0$  if  $\kappa = 1$  (the non-symmetric scheme) and in general we have,

$$\begin{aligned} |T_{10}| &\leq 2(1-\kappa)k \sum_{i=1}^m \sum_{e \in \Gamma_h \cup \Gamma_D} \left(\frac{|e|^\beta}{\delta}\right)^{1/2} \|\{D\nabla \bar{\chi}_i \cdot \mathbf{n}_e\}\|_{0,e} \left(\frac{\delta}{|e|^\beta}\right)^{1/2} \|\bar{\chi}_i\|_{0,e}, \\ &\leq 2(1-\kappa)k \sum_{i=1}^m \frac{Ch^{(d-1)\beta/2-1/2}}{\delta^{1/2}} \|D^{1/2} \nabla \bar{\chi}_i\|_0 J_0^{\delta,\beta}(\bar{\chi}_i, \bar{\chi}_i)^{1/2}, \\ &\leq (1-\kappa)\epsilon_{10}k \sum_{i=1}^m \frac{Ch^{(d-1)\beta-1}}{\delta} \|\bar{\chi}_i\|_{\mathcal{A}}^2 + \frac{(1-\kappa)k}{\epsilon_{10}} \sum_{i=1}^m \|\bar{\chi}_i\|_{\mathcal{A}}^2. \end{aligned}$$

For  $T_{11}$  a similar argument produces,

$$|T_{11}| \leq \epsilon_{11}k \sum_{i=1}^m \frac{Ch^{(d-1)\beta-1}}{\delta} \|\bar{\boldsymbol{\eta}}_i\|_0^2 + \frac{k}{\epsilon_{11}} \sum_{i=1}^m \|\bar{\chi}_i\|_{\mathcal{A}}^2,$$

and, for  $T_{12}$ ,

$$\begin{aligned} |T_{12}| &\leq \epsilon_{12}k \sum_{i=1}^m \frac{Ch^{(d-1)\beta}}{\delta} \left(\sum_E \|\bar{\boldsymbol{\theta}}_i\|_{L_2(\partial E)}\right)^2 + \frac{k}{\epsilon_{12}} \sum_{i=1}^m J_0^{\delta,\beta}(\bar{\chi}_i, \bar{\chi}_i), \\ &\leq \frac{Ct_m \epsilon_{12} h^{2r-1+(d-1)\beta}}{\delta} \|\boldsymbol{\sigma}\|_{L_\infty(0,t_m;H^r(\Omega))}^2 + \frac{k}{\epsilon_{12}} \sum_{i=1}^m J_0^{\delta,\beta}(\bar{\chi}_i, \bar{\chi}_i). \end{aligned}$$

Setting  $\epsilon_{10} = 2$ , and choosing

$$\epsilon_3 + \epsilon_5 + \epsilon_8 + \frac{1}{\epsilon_{12}} = \frac{1}{4}, \quad \epsilon_4 + \epsilon_7 + \epsilon_9 = 1 \quad \text{and} \quad \epsilon_{11} = 4,$$

we then assemble these estimates and obtain,

$$\begin{aligned} \|\chi_m\|_0^2 + \frac{K}{\mu} \|\boldsymbol{\eta}_m\|_0^2 + \left(\frac{1}{2} - \frac{4C\hat{h}^{(d-1)\beta-1}}{\delta}\right) k \sum_{i=1}^m \|\bar{\chi}_i\|_{\mathcal{A}}^2 \\ + \left(1 - \frac{4\mu C\hat{h}^{(d-1)\beta-1}}{\delta K\gamma_G}\right) k \sum_{i=1}^m \frac{K\gamma_G}{\mu} \|\bar{\boldsymbol{\eta}}_i\|_0^2 \\ \leq C(h^{2r} + k^4) + \frac{2kK}{\mu} \sum_{i=1}^m \left| \left(\overline{\gamma(u)\boldsymbol{\sigma}(t_i)} - \gamma(\mathcal{B}_{i,n} u^h) \bar{\boldsymbol{\sigma}}_i^*, \bar{\boldsymbol{\eta}}_i\right) \right|, \end{aligned}$$

where we recalled that  $\beta \geq (d-1)^{-1}$ . Now we make several appeals to Lemma 3.3. Firstly, when  $n = 1$  we have, for  $k \leq \hat{k}$ , that,

$$\begin{aligned} & \left(1 - \frac{C\hat{k}}{\epsilon}\right) \|\chi_m\|_0^2 + \frac{K}{\mu} \|\boldsymbol{\eta}_m\|_0^2 + \left(\frac{1}{2} - \frac{4C\hat{h}^{(d-1)\beta-1}}{\delta}\right) k \sum_{i=1}^m \|\bar{\chi}_i\|_{\mathcal{A}}^2 \\ & + \left(1 - \frac{4\mu C\hat{h}^{(d-1)\beta-1}}{\delta K \gamma_G} - \epsilon\right) k \sum_{i=1}^m \frac{K\gamma_G}{\mu} \|\bar{\boldsymbol{\eta}}_i\|_0^2 \leq C(h^{2r} + k^4) + \frac{Ck}{\epsilon} \sum_{i=0}^{m-1} \|\chi_i\|_0^2. \end{aligned}$$

Choosing  $\epsilon = 1/2$ ,  $\hat{k}$  and  $\hat{h}^{(d-1)\beta-1}/\delta$  small enough, an application of Gronwall's lemma then results in,

$$\|\chi_m\|_0^2 + \|\boldsymbol{\eta}_m\|_0^2 + k \sum_{i=1}^m \|\bar{\chi}_i\|_{\mathcal{A}}^2 + k \sum_{i=1}^m \|\bar{\boldsymbol{\eta}}_i\|_0^2 \leq C(h^{2r} + k^4).$$

Secondly, for the linearised scheme where  $n = 2$ , we have, by Lemma 3.3 for  $m = 1$ , and with  $\epsilon = (2\gamma_G k)^{-1}$  that,

$$\begin{aligned} & \|\chi_1\|_0^2 + \frac{K}{2\mu} \|\boldsymbol{\eta}_1\|_0^2 + \left(\frac{1}{2} - \frac{4C\hat{h}^{(d-1)\beta-1}}{\delta}\right) k \|\bar{\chi}_1\|_{\mathcal{A}}^2 \\ & + \left(1 - \frac{4\mu C\hat{h}^{(d-1)\beta-1}}{\delta K \gamma_G}\right) k \frac{K\gamma_G}{\mu} \|\bar{\boldsymbol{\eta}}_1\|_0^2 \leq C(h^{2r} + k^4) + Ck^2 \|\chi_0\|_0^2 + C\|\boldsymbol{\eta}_0\|_0^2. \end{aligned}$$

Now use the estimates given above for  $T_1$  and  $T_2$  and again select  $\hat{h}^{(d-1)\beta-1}/\delta$  small enough to get,

$$\|\chi_1\|_0^2 + \|\boldsymbol{\eta}_1\|_0^2 + k \|\bar{\chi}_1\|_{\mathcal{A}}^2 + k \|\bar{\boldsymbol{\eta}}_1\|_0^2 \leq C(h^{2r} + k^4).$$

On the other hand, for  $m > 1$  we estimate the first term in the sum (corresponding to  $i = 1$ ) in  $T_{13}$  by choosing  $\epsilon = 1/k$  in Lemma 3.3 and then use the estimates just obtained. For the remaining terms we choose  $\epsilon = 1/2$ . With empty sums set to zero, we then have for  $m = 2, 3, 4, \dots$  that,

$$\begin{aligned} & \|\chi_m\|_0^2 + \frac{K}{\mu} \|\boldsymbol{\eta}_m\|_0^2 + \left(\frac{1}{2} - \frac{4C\hat{h}^{(d-1)\beta-1}}{\delta}\right) k \sum_{i=1}^m \|\bar{\chi}_i\|_{\mathcal{A}}^2 \\ & + \left(\frac{1}{2} - \frac{4\mu C\hat{h}^{(d-1)\beta-1}}{\delta K \gamma_G}\right) k \sum_{i=1}^m \frac{K\gamma_G}{\mu} \|\bar{\boldsymbol{\eta}}_i\|_0^2 \leq C(h^{2r} + k^4) + Ck \sum_{i=2}^{m-1} \|\chi_i\|_0^2, \end{aligned}$$

by the same estimates for the initial conditions as used previously. Once again, we choose  $\hat{h}^{(d-1)\beta-1}/\delta$  small enough and use Gronwall's lemma to arrive at,

$$\|\chi_m\|_0^2 + \|\boldsymbol{\eta}_m\|_0^2 + k \sum_{i=1}^m \|\bar{\chi}_i\|_{\mathcal{A}}^2 + k \sum_{i=1}^m \|\bar{\boldsymbol{\eta}}_i\|_0^2 \leq C(h^{2r} + k^4).$$

We now see that this inequality holds for all  $m \in \{1, \dots, N\}$  in both of the cases  $n = 1$  and  $n = 2$ . By the triangle inequality we then have,

$$\begin{aligned} & \|u(t_m) - u_m^h\|_0 + \|\boldsymbol{\sigma}(t_m) - \boldsymbol{\sigma}_i^h\|_0 + \left(k \sum_{i=1}^m \|\bar{u}(t_i) - \bar{u}_i^h\|_{\mathcal{A}}^2\right)^{1/2} + \left(k \sum_{i=1}^m \|\bar{\boldsymbol{\sigma}}(t_i) - \bar{\boldsymbol{\sigma}}_i^h\|_0^2\right)^{1/2} \\ & \leq \|\xi(t_m)\|_0 + \|\boldsymbol{\theta}(t_m)\|_0 + \left(k \sum_{i=1}^m \|\bar{\xi}(t_i)\|_{\mathcal{A}}^2\right)^{1/2} + \left(k \sum_{i=1}^m \|\bar{\boldsymbol{\theta}}(t_i)\|_0^2\right)^{1/2} \\ & + \|\chi(t_m)\|_0 + \|\boldsymbol{\eta}(t_m)\|_0 + \left(k \sum_{i=1}^m \|\bar{\chi}(t_i)\|_{\mathcal{A}}^2\right)^{1/2} + \left(k \sum_{i=1}^m \|\bar{\boldsymbol{\eta}}(t_i)\|_0^2\right)^{1/2}, \end{aligned}$$

and our estimates, along with (16) and (30) and the fact that  $(a^2 + b^2)^{1/2} \leq a + b$  for  $a, b \geq 0$ , then complete the proof.  $\square$

Note that due to the much larger number of terms involved this proof used Gronwall's inequality, unlike the proof of Theorem 2.2. It is possible that more careful estimation could remove the need for an exponentially large 'Gronwall constant' in the error estimate, but we leave this as a problem for another time.

If we replace the  $\mathcal{D}_r(\mathcal{E}_h)$ -approximation of  $u$  by a standard conforming piecewise polynomial finite element space containing the essential boundary condition, then the DG FEM schemes presented above reduce to a standard continuous Galerkin (CG) FEM. An error estimate of the form presented in Theorem 3.4 then continues to hold (as a special case).

**Corollary 3.5** *For a CG finite element approximation of the problem we also have,*

$$\|u(t_m) - u_m^h\|_0 + \|\boldsymbol{\sigma}(t_m) - \boldsymbol{\sigma}_m^h\|_0 + \left( k \sum_{i=1}^m \|\bar{u}_i - \bar{u}_i^h\|_{\mathcal{A}}^2 + \|\bar{\boldsymbol{\sigma}}_i - \bar{\boldsymbol{\sigma}}_i^h\|_0^2 \right)^{1/2} \leq C(h^r + k^2)$$

for each  $m = 1, \dots, N$ .

**Remark 3.6** *If  $(u, \boldsymbol{\sigma})$  is a solution of (12), (13), then we could use Theorems 1.1 and 3.4 to show that,*

$$\|u_m^h\|_0^2 + \|\boldsymbol{\sigma}_m^h\|_0^2 \leq C(u).$$

*This would follow from the triangle inequality and is the closest we can get to a stability estimate. However, to get 'data' on the right hand side we would need stability estimates on higher derivatives of the exact solutions.*

Theorem 3.4 naturally contains some regularity assumptions on both  $u$  and  $\boldsymbol{\sigma}$ . Since, via (10), we can replace the system (1) and (2) by the single (11) we can expect that the regularity of  $\boldsymbol{\sigma}$  can be tied into that of  $u$ . In this direction, for the case of piecewise linear spatial approximation ( $r = 1$ ), we have the following claim (see [18] for details).

**Proposition 3.7** *For  $r = 1$  the regularity requirements of Theorem 3.4 can be replaced by,*

$$u \in H^2(I; H^2(\Omega)) \cap L_1(I; W_\infty^1(\Omega)) \cap L_\infty(I; W_4^1(\Omega)) \cap W_8^1(I; L_8(\Omega)) \\ \cap W_4^2(I; L_4(\Omega)) \cap H^3(I; L_2(\Omega)) \cap H^1(I; W_4^1(\Omega))$$

and  $\boldsymbol{\sigma} \in L_\infty(\Omega) \cap \mathbf{H}^1(\Omega)$ .

## 4 Numerical experiments

We anticipate that the linearised scheme is the one that is of most interest and so first quote from [4] just a few numerical results to illustrate Theorem 3.4 in the case  $r = 1$ . The data common to these first results are:  $D = K = \mu = 1$ ,  $\gamma_R = 10$ ,  $\gamma_G = 0.1$ ,  $\Delta = 0.1$ ,  $\Omega = (0, 1)^2$  and  $I = (0, T)$  for  $T = 1$ , and in each case the loads and boundary conditions are designed so that the problem has a known exact solution. (To achieve this we added a function  $\mathbf{h} = \mathbf{h}(\mathbf{x}, t)$  to the right of (2).) The resulting errors,

$$\mathcal{E} := k \sum_{i=1}^N \left( \|\bar{u}_i - \bar{u}_i^h\|_{\mathcal{A}}^2 + \|\bar{\boldsymbol{\sigma}}_i - \bar{\boldsymbol{\sigma}}_i^h\|_0^2 \right)^{1/2}$$

are tabulated along with the estimated order of convergence (EOC). In the tables,  $M$  denotes a uniform  $M \times M$  space mesh and  $N$  is the number of time intervals.

$M$	$\kappa = -1$		$\kappa = 1$	
	$\mathcal{E}$	EOC	$\mathcal{E}$	EOC
1	2.032013		2.026638	
2	1.837297	0.1453	1.837534	0.1413
4	1.532320	0.2619	1.532354	0.2620
8	0.863881	0.8268	0.863874	0.8269
16	0.447608	0.9486	0.447608	0.9486
32	0.225955	0.9862	0.225955	0.9862

Table 1: Tabulated errors for  $N = 2$ ,  $\delta = 10^2$  and  $\beta = 3$ .

$N$	$\kappa = -1$		$\kappa = 1$	
	$\mathcal{E}$	EOC	$\mathcal{E}$	EOC
1	$8.845824 \times 10^{-2}$		$8.845825 \times 10^{-2}$	
2	$2.561600 \times 10^{-2}$	1.7879	$2.561600 \times 10^{-2}$	1.7879
4	$6.598674 \times 10^{-3}$	1.9568	$6.598677 \times 10^{-3}$	1.9568
8	$1.660746 \times 10^{-3}$	1.9903	$1.660747 \times 10^{-3}$	1.9903
16	$4.166057 \times 10^{-4}$	1.9951	$4.166061 \times 10^{-4}$	1.9951
32	$1.049564 \times 10^{-4}$	1.9889	$1.049566 \times 10^{-4}$	1.9889
64	$2.707173 \times 10^{-5}$	1.9549	$2.707180 \times 10^{-5}$	1.9549

Table 2: Tabulated errors for  $M = 8$ ,  $\delta = 10^4$  and  $\beta = 2$ .

Table 1 shows results for the solutions,

$$u(\mathbf{x}, t) = t \sin(2\pi x) \sin(2\pi y), \quad \boldsymbol{\sigma}(\mathbf{x}, t) = t \begin{pmatrix} \sin(2\pi x) \\ \cos(2\pi y) \end{pmatrix},$$

in the case  $\Gamma_D = \{x = 0 \text{ or } y = 0\}$  when  $u_{RG} = 0.5$ . In this case there is no time discretisation error and we observe  $O(h)$  convergence.

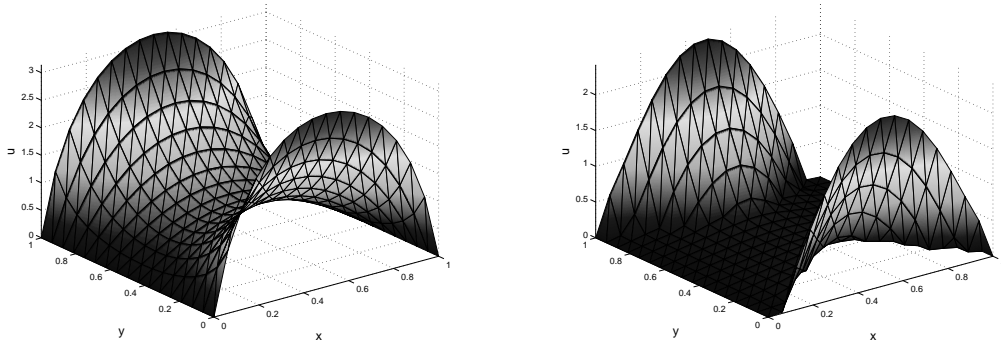


Figure 1: Computed surfaces showing  $u$  at  $t = 10$  for the data  $\Omega := (0, 1)^2$ ,  $f = 0$ ,  $g = 1$ ,  $\Gamma_N := \{y = 0 \text{ or } y = 1\}$ ,  $\check{u} = -0.5x(x - 1)$ ,  $\check{\boldsymbol{\sigma}} = \mathbf{0}$ ,  $D = 10^{-1}$ ,  $K = 10^{-4}$ ,  $\mu = 10^4$ ,  $\Delta = 10^{-3}$  and  $u_{RG} = 0.5$ . The figure on the left corresponds to  $\gamma_R = \gamma_G = 5000$  and the one on the right to  $\gamma_R = 10^4$  and  $\gamma_G = 10^{-3}$ .

On the other hand, for the solutions

$$u(\mathbf{x}, t) = t^3 x \quad \boldsymbol{\sigma}(\mathbf{x}, t) = t^2 \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

there is no space discretisation error and for  $\Gamma_D = \{x = 0\}$  with  $u_{RG} = 0.5$  we observe, in Table 2,  $O(k^2)$  convergence.

Cohen *et al.*'s model, [6], produces solutions which exhibit very sharp changes in  $u$ , and these fronts become steeper as time advances. It seems that the near-discontinuity in  $u$  is driven by the fact that their scalar stress equation is,

$$\sigma_t + \gamma(u)\sigma = \mu u,$$

whereas ours is a vector equation given by (2) and has the gradient of  $u$  on the right. Because of this, solutions to our model exhibit sharp changes in  $\nabla u$  rather than  $u$  itself, and we illustrate this in Figure 1 ( $16 \times 16$  elements, 50 time steps,  $\beta = 2$ ,  $\delta = 10^2$ ,  $\kappa = -1$ ).

The surface plot on the left corresponds to linear non-Fickian behaviour where we choose  $\gamma_G = \gamma_R$  so as to remove the nonlinear term in (5). The figure on the right shows the effect of the nonlinearity when  $\gamma_R \gg \gamma_G$ , and we can see steep changes in  $\nabla u$ .

## 5 Conclusion

The numerical experiments support the error estimate in Theorem 3.4 and so we conclude that the linearisation derived from the extrapolation is an effective method of approximating the solution to this type of problem. Also, on examining the estimates in Lemma 3.3, we see that the linearised scheme does not require any additional regularity assumptions. Hence, we conclude that it should always be preferred over the nonlinear scheme.

As we mentioned earlier in Section 1, our model is a simplification of the original model proposed in [6]. Nonetheless, preliminary numerical experiments (not included here) with CG FEM indicate that it is capable of capturing the same basic phenomena of steep travelling fronts. An error analysis for the model in [6] is currently being undertaken and, when complete, we expect to give more extensive numerical demonstrations for both models.

On a closing note, the problem we have studied is a generalisation of a parabolic analogue to the dynamic solids problem considered in [20] to the case of nonlinear relaxation time. It is an ongoing project to extend our results to the dynamic case, and to other types of nonlinearities (see for example the nonlinear relaxation time discussed in [22]).

## References

- [1] Robert A. Adams and John J. F. Fournier. *Sobolev spaces*, volume 140 of *Pure and Applied Mathematics*. Academic Press, second edition, 2003.
- [2] Herbert Amann. Global existence for a class of highly degenerate parabolic systems. *Japan J. Indust. Appl. Math.*, 8:143—151, 1991.
- [3] H.T. Banks, Gabriella A. Pintér, and Laura K. Potter. Existence of unique weak solutions to a dynamical system for nonlinear elastomers with hysteresis. Technical Report CRSC-TR98-43, Center for Research in Scientific Computation, North Carolina State University, 1998. [www.ncsu.edu/crsc/reports.htm/](http://www.ncsu.edu/crsc/reports.htm/).
- [4] Roswitha Bultmann and Simon Shaw. Finite element discretisation of a problem in nonlinear non-Fickian viscoelastic diffusion using a discontinuous Galerkin method in space. Technical report, BICOM, Brunel University, 2005. Technical Report 05/5, [www.brunel.ac.uk/bicom](http://www.brunel.ac.uk/bicom).

- [5] J. R. Cannon and Y. Lin. *A priori  $L^2$  error estimates for finite-element methods for nonlinear diffusion equations with memory.* *SIAM J. Numer. Anal.*, 27:595—607, 1990.
- [6] D. S. Cohen, A. B. White Jr., and T. P. Witelski. Shock formation in a multidimensional viscoelastic diffusive system. *SIAM J. Appl. Math.*, 55:348—368, 1995.
- [7] J. Douglas, R.E. Ewing, and M.F. Wheeler. A time-discretization procedure for a mixed finite element approximation of miscible displacement in porous media. *RAIRO Numerical Analysis*, 17(3):249—265, 1983.
- [8] J. D. Ferry. *Viscoelastic properties of polymers.* John Wiley and Sons Inc., 1970.
- [9] V. Girault, B. Riviere, and M. Wheeler. A discontinuous Galerkin method with non-overlapping domain decomposition for the Stokes and Navier-Stokes problems. *Mathematics of Computation*, TICAM Report 02-08 (2002), to appear.
- [10] A. R. Johnson. Modeling viscoelastic materials using internal variables. *The Shock and Vibration Digest*, 31:91—100, 1999.
- [11] A. R. Johnson, A. Tessler, and M. Dambach. Dynamics of thick viscoelastic beams. *Journal of Engineering Materials and Technology*, 119:273—278, 1997.
- [12] CH. Lubich, I. H. Sloan, and V. Thomée. Nonsmooth data error estimates for approximations of an evolution equation with a positive-type memory term. *Math. Comp.*, 65:1—17, 1996.
- [13] W. Mclean and V. Thomée. Numerical solution of an evolution equation with a positive-type memory term. *J. Austral. Math. Soc.*, 35:23—70, 1993.
- [14] J. W. Nunziato. On heat conduction in materials with memory. *Quart. Appl. Maths.*, 29:187—204, 1971.
- [15] B. Rivière, M. F. Wheeler, and V. Girault. A priori error estimates for finite element methods based on discontinuous approximation spaces for elliptic problems. *SIAM J. Numer. Anal.*, 39:902–931, 2001.
- [16] B. Rivière and M.F. Wheeler. A discontinuous Galerkin method applied to nonlinear parabolic equations. In B. Cockburn, G.E. Karniadakis, and C.-W. Shu, editors, *Discontinuous Galerkin Methods: Theory, Computation and Applications*, volume 11 of *Lecture Notes in Computational Science and Engineering*, pages 231–244. Springer, 1999.
- [17] B. Riviere and M.F. Wheeler. Discontinuous Galerkin methods for flow and transport problems in porous media. *Communications in Numerical Methods in Engineering*, 18:63—68, 2002.
- [18] Béatrice Rivière and Simon Shaw. Discontinuous Galerkin finite element approximation of nonlinear non-Fickian diffusion in viscoelastic polymers. Technical report, BICOM, Brunel University, 2004. Technical Report 04/3, [www.brunel.ac.uk/bicom](http://www.brunel.ac.uk/bicom).
- [19] Béatrice Rivière, Simon Shaw, Mary F. Wheeler, and J.R. Whiteman. Discontinuous Galerkin finite element methods for linear elasticity and quasistatic linear viscoelasticity. *Numer. Math.*, 95:347—376, 2003.
- [20] Béatrice Rivière, Simon Shaw, and J.R. Whiteman. Discontinuous Galerkin finite element methods for dynamic linear solid viscoelasticity problems. Submitted to *Numer. Methods Partial Differential Equations* (see also report 05/7 at [www.brunel.ac.uk/bicom](http://www.brunel.ac.uk/bicom)).

- [21] Béatrice Rivière, Mary F. Wheeler, and Vivette Girault. Improved energy estimates for interior penalty, constrained and discontinuous Galerkin methods for elliptic problems. Part I. *Computational Geosciences*, 3:337—360, 1999.
- [22] Simon Shaw, M. K. Warby, and J. R. Whiteman. Numerical techniques for problems of quasistatic and dynamic viscoelasticity. In J. R. Whiteman, editor, *The Mathematics of Finite Elements and Applications*. MAFELAP 1993, pages 45—68. Wiley, Chichester, 1994.
- [23] I. H. Sloan and V. Thomée. Time discretization of an integro-differential equation of parabolic type. *SIAM J. Numer. Anal.*, 23:1052—1061, 1986.
- [24] N. L. Thomas and A. H. Windle. A theory of Case II diffusion. *Polymer*, 23:529—542, 1982.
- [25] Noreen Thomas and A. H. Windle. Transport of methanol in poly(methyl-methacrylate). *Polymer*, 19:255—265, 1978.
- [26] V. Thomée and L. B. Wahlbin. Long-time numerical solution of a parabolic equation with memory. *Math. Comp.*, 62:477—496, 1994.