FINITE ELEMENT CONVERGENCE ANALYSIS FOR THE THERMOVISCOELASTIC JOULE HEATING PROBLEM

AXEL MÅLQVIST AND TONY STILLFJORD

Abstract. We consider a system of equations that model the temperature, electric potential and deformation of a thermoviscoelastic body. A typical application is a thermistor; an electrical component that can be used e.g. as a surge protector, temperature sensor or for very precise positioning. We introduce a full discretization based on standard finite elements in space and a semi-implicit Euler-type method in time. For this method we prove optimal convergence orders, i.e. second-order in space and first-order in time. The theoretical results are verified by several numerical experiments in two and three dimensions.

1. Introduction

Consider the following system of coupled equations:

$$\dot{\theta} = \Delta\theta + \sigma(\theta)|\nabla\phi|^2 - \mathbf{M} : \epsilon(\dot{u}), \tag{1.1}$$

$$0 = \nabla \cdot \big(\sigma(\theta)\nabla\phi\big), \tag{1.2}$$

$$\ddot{u} = \nabla \cdot \big(\mathbf{A}\epsilon(\dot{u}) + \mathbf{B}\epsilon(u) - \mathbf{M}\theta\big) + f, \tag{1.3}$$

with initial conditions

$$\theta(0,x) = \theta_0(x), \quad u(0,x) = u_0(x) \quad \text{and} \quad \dot{u}(0,x) = v_0(x),$$

over the convex polygonal or polyhedral domain $\Omega \subset \mathbb{R}^d$ with $d \leq 3$. Together with appropriate boundary conditions, to be specified later, these equations describe the evolution of the temperature θ, electric potential ϕ and deformation u of a conducting body. Here $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{M}$ are constant tensors, describing the viscosity, elasticity and thermal expansion of the body. The vector f consists of external forces and $\sigma(\theta)$ denotes the electrical conductivity, which here depends on the temperature. In addition, we have used the notation

$$\epsilon(u) = \frac{1}{2}\big(\nabla u + (\nabla u)^T\big)$$

for the linearized strain tensor and : for the Frobenius inner product.

The coupling of electricity and temperature through (1.1)–(1.2) is commonly known as *Joule heating* and is typically used to model thermistors, see e.g. [5, 8]. These are electrical components used for example as surge protectors or temperature sensors. The inclusion of thermoviscoelastic effects through (1.3) allows us to also model their use as actuators on the micro-scale, cf. [15].

Date: November 2, 2016.

2010 *Mathematics Subject Classification.* 65M12, 65M60, 74D05, 74H15.

Key words and phrases. Partial differential equations, thermoviscoelastic, Joule heating, thermistor, convergence analysis, finite elements.

Both authors were supported by the Swedish Research Council under grant 2015-04964.

We note that the Joule heating problem, both stationary and time-dependent, has been considered extensively in different contexts. For discussions on existence and uniqueness, see e.g. [2, 5, 6, 7, 8, 16, 17, 18, 27] and the references therein. For the fully coupled, deformable problem the literature is less extensive. We refer mainly to [19] for the non-degenerate case that we consider here, with $\sigma \geq \sigma_{\min} > 0$. See also [26] for the degenerate case where $\sigma = 0$ is allowed; this requires a more generalized solution concept.

However, to our knowledge there exists no numerical analysis for methods applied to the fully coupled case. Many authors have analyzed methods for similar problems. For example, [11] considers the quasi-static version where the $\ddot{u}$-term is ignored, [1], [10] and [20] considers the non-deformable case, [12, 13] treat the purely thermoviscoelastic case (no ϕ) with nonlinear constituent law, etc. Additionally, in the deformable case a common theme seems to be suboptimal convergence orders, i.e. errors of the form $\mathrm{O}(h+k)$ instead of $\mathrm{O}(h^2+k)$.

The main contribution of this article is therefore an error analysis for a fully discrete discretization applied to the problem (1.1)–(1.3), which shows optimal convergence orders in both time and space. For the spatial discretization we consider standard finite elements, and for the temporal discretization a semi-implicit Euler-type method. Our approach also allows us to analyze e.g. the implicit Euler method, but the semi-implicit method benefits from a greatly decreased computational cost while the errors are comparable.

The central idea of our proof is to bound the errors in ϕ and $\dot{u}$ in terms of the error in θ, in the spirit of [10] and [21]. The latter error then fulfills an equation similar to (1.1), to which we may apply a Grönwall inequality after properly handling the quadratic potential term. We note that we avoid any time step restrictions of the form $k \leq h^{d/r}$ by performing the analysis in two steps, where the first considers only the discretization in time, cf. [21]. Finally, in order to produce the $\dot{u}$ error bound, we extend the concept of Ritz-Volterra projections for damped wave equations (see [22]) to the discrete and vector-valued viscoelasticity case.

For simplicity, we consider Dirichlet boundary conditions,

$$\theta(t,x) = 0, \quad \phi(t,x) = \phi_b(t,x) \quad \text{and} \quad u(t,x) = 0$$

for $t \in [0,T]$ and $x \in \partial\Omega$. This is a simplified case of the ideal situation with an arbitrary polygon and mixed boundary conditions, corresponding to where the body is clamped and insulated. As is well known (see e.g. [14]) the solutions to such a problem would typically suffer from a lack of regularity in the vicinity of re-entrant corners and boundary condition transitions, which leads to suboptimal convergence orders for finite-element based numerical methods. We therefore restrict ourselves to the simplified model, and will indicate possible generalizations by our numerical experiments.

A brief outline of the article is as follows. In Section 2 we write the problem on weak form and discretize it in both time and space. The assumptions on the data and solutions to the continuous problem are given in Section 3, where we also perform the error analysis. In Subsection 3.1, the time-discrete system is shown to be first-order convergent, and then the full discretization is shown to be second-order convergent to the time-discrete system in Subsection 3.2. These results are confirmed by the numerical experiments presented in Section 4, and conclusions and future work is summarized in Section 5.

2. Weak formulation and discretization

In order to present a weak formulation of the problem, we introduce the spaces

$$V := H_0^1(\Omega) \subset L^2(\Omega), \quad \text{and} \quad \boldsymbol{V} := H_0^1(\Omega)^d \subset L^2(\Omega)^d =: \boldsymbol{L}^2(\Omega)$$

as well as the space of second-order symmetric tensors,

$$Q = \{\xi = (\xi_{ij})_{i,j=1}^d \subset L^2(\Omega)^{d\times d}\ ;\ \xi_{ji} = \xi_{ij}, 1 \le i,j \le d\}.$$

The idea here is that θ and $\phi - \phi_b$ belong to V, $u \in \boldsymbol{V}$ and $\epsilon(u) \in Q$. On Q, we have the inner product

$$(\xi, \zeta)_Q := \int_\Omega \xi(x) : \zeta(x) \,\mathrm{d}x = \sum_{i,j=1}^d (\xi_{ij}, \zeta_{ij})_{L^2(\Omega)}.$$

which gives rise to the norm $\|\cdot\|_Q$. To simplify some notation, we use the inner product

$$(u, v)_{\boldsymbol{V}} = (\epsilon(u), \epsilon(v))_Q$$

on $\boldsymbol{V}$ instead of the usual one. The norm $\|\cdot\|_{\boldsymbol{V}}$ induced by this inner product is equivalent to $\|\cdot\|_{H^1(\Omega)^d}$ by Korn's inequality, see e.g. [9, Chapter III, Theorems 3.1, 3.3] and [24]. We will on several occasions make use also of the norm $\|\cdot\|_{\mathbf{B}}$, which arises from the elasticity operator through

$$\|u\|_{\mathbf{B}}^2 = (\mathbf{B}\epsilon(u), \epsilon(u)),$$

as well as the norm $\|\cdot\|_{\mathbf{A}+k\mathbf{B}}$ defined analogously for a small positive constant k. Under Assumption 3.1 in the next section, both of these norms are equivalent to the $\boldsymbol{V}$-norm. In the following, we will omit the specification of Ω and simply write L^2 or $\boldsymbol{L}^2$. Additionally, the L^2- and $\boldsymbol{L}^2$-norms will both simply be denoted by $\|\cdot\|$ and the corresponding inner products by $(\cdot,\cdot)$, where no confusion can arise.

By multiplying the equations (1.1), (1.2) with the test function $\chi \in V$, Equation (1.3) with $\boldsymbol{\chi} \in \boldsymbol{V}$ and then using Green's formula combined with the identity $(\epsilon(u), \nabla v) = (\epsilon(u), \epsilon(v))$, we get

$$\left(\dot\theta, \chi\right) + (\nabla\theta, \nabla\chi) = \left(\sigma(\theta)|\nabla\phi|^2, \chi\right) - (\mathbf{M} : \epsilon(\dot u), \chi), \tag{2.1}$$

$$(\sigma(\theta)\nabla\phi, \nabla\chi) = 0, \tag{2.2}$$

$$(\ddot u, \boldsymbol{\chi}) + (\mathbf{A}\epsilon(\dot u) + \mathbf{B}\epsilon(u), \epsilon(\boldsymbol{\chi}))_Q = (\mathbf{M}\theta, \epsilon(\chi))_Q + (f, \boldsymbol{\chi}), \tag{2.3}$$

for all $\chi \in V$ and $\boldsymbol{\chi} \in \boldsymbol{V}$, respectively. Note that we have omitted the time parameter here and in the original equation; both are supposed to hold for all times $t \in (0, T]$ for a given T.

We now discretize the time interval $[0, T]$ using a constant temporal step size k, which results in the grid $t_n = nk$ with $n = 1, 2, \ldots, N$ and $Nk = T$. We will abbreviate function evaluations at these times by sub-scripts, so that

$$\theta_n = \theta(t_n), \quad \phi_n = \phi(t_n), \quad u_n = u(t_n) \quad \text{and} \quad f_n = f(t_n).$$

The approximations of these solution values should belong to the same spaces as in the continuous case, and we will denote them by capital letters and superscripts:

$$\Theta^n \approx \theta_n, \quad \Phi^n \approx \phi_n \quad \text{and} \quad U^n \approx u_n.$$

Additionally, we denote by $\mathrm{D_t}$ the first-order backward difference quotient, i.e.

$$\mathrm{D_t}\,\Theta^n = \frac{\Theta^n - \Theta^{n-1}}{k}.$$

With this notation given, we now consider the following semi-implicit temporal discretization of Equations (1.1)–(1.3),

$$\mathrm{D_t}\,\Theta^n = \Delta\Theta^n + \sigma(\Theta^{n-1})|\nabla\Phi^{n-1}|^2 - \mathbf{M} : \epsilon(\mathrm{D_t}\,U^{n-1}), \tag{2.4}$$

$$0 = \nabla\cdot\big(\sigma(\Theta^n)\nabla\Phi^n\big), \tag{2.5}$$

$$\mathrm{D_t^2}\,U^n = \nabla\cdot\big(\mathbf{A}\epsilon(\mathrm{D_t}\,U^n) + \mathbf{B}\epsilon(U^n) - \mathbf{M}\Theta^n\big) + f_n, \tag{2.6}$$

and its corresponding weak form,

$$(\mathrm{D_t}\,\Theta^n, \chi) + (\nabla\Theta^n, \nabla\chi) = \big(\sigma(\Theta^{n-1})|\nabla\Phi^{n-1}|^2, \chi\big) - \big(\mathbf{M} : \epsilon(\mathrm{D_t}\,U^{n-1}), \chi\big)\,, \tag{2.7}$$

$$(\sigma(\Theta^n)\nabla\Phi^n, \nabla\chi) = 0, \tag{2.8}$$

$$\big(\mathrm{D_t^2}\,U^n, \boldsymbol{\chi}\big) + (\mathbf{A}\epsilon(\mathrm{D_t}\,U^n) + \mathbf{B}\epsilon(U^n), \epsilon(\boldsymbol{\chi}))_Q = (\mathbf{M}\Theta^n, \epsilon(\chi))_Q + (f_n, \boldsymbol{\chi})\,, \tag{2.9}$$

for $n = 1, \ldots, N$ and for all $\chi \in S_h$ and $\boldsymbol{\chi} \in \boldsymbol{S}_h$, respectively. The initial conditions are the same as in the continuous case: $\Theta^0 = \theta_0$, $U^0 = u_0$ and $\mathrm{D_t}\,U^1 = v_0$. Note that this discretization results in a decoupling of the equations; we solve first for Θ^n using (2.4) then use this to find Φ^n from (2.5) and U^n from (2.6). This implies a significant decrease in computational effort compared to the fully coupled case arising from e.g. the implicit Euler discretization.

For the spatial discretization, we introduce the finite element spaces $S_h \subset V$ and $\boldsymbol{S}_h \subset \boldsymbol{V}$. These consist of continuous, piecewise linear functions with zero trace on $\partial\Omega$, defined on a quasi-uniform mesh with mesh-width h. Then the fully discrete problem we are interested in is given by

$$(\mathrm{D_t}\,\Theta_h^n, \chi) + (\nabla\Theta_h^n, \nabla\chi) = \big(\sigma(\Theta_h^{n-1})|\nabla\Phi_h^{n-1}|^2, \chi\big) - \big(\mathbf{M} : \epsilon(\mathrm{D_t}\,U_h^{n-1}), \chi\big)\,, \tag{2.10}$$

$$(\sigma(\Theta_h^n)\nabla\Phi_h^n, \nabla\chi) = 0, \tag{2.11}$$

$$\big(\mathrm{D_t^2}\,U_h^n, \boldsymbol{\chi}\big) + (\mathbf{A}\epsilon(\mathrm{D_t}\,U_h^n) + \mathbf{B}\epsilon(U_h^n), \epsilon(\boldsymbol{\chi}))_Q = (\mathbf{M}\Theta_h^n, \epsilon(\chi))_Q + (f_n, \boldsymbol{\chi})\,, \tag{2.12}$$

for $n = 1, \ldots, N$ and for all $\chi \in S_h$ and $\boldsymbol{\chi} \in \boldsymbol{S}_h$, respectively. Here, the approximations satisfy $\Theta_h^n \in S_h$, $\Phi_h^n - \phi_b(t_n) \in S_h$ and $U_h^n \in \boldsymbol{S}_h$. As initial conditions, we take $U_h^0 = 0$, $\mathrm{D_t}\,U_h^1 = U_h^1 = 0$ and $\Theta_h^0 = I_h\theta_0$, the Lagrangian interpolant of the exact initial condition.

3. Error analysis

Our main goal is to estimate the errors $\|\Theta_h^n - \theta_n\|$, $\|\Phi_h^n - \phi_n\|$ and $\|U_h^n - u_n\|$. In order to do this, we will generalize the analysis of [21] (cf. also [10]) for the case with no deformation. This consists of first showing that the time-discrete approximations are $\mathrm{O}(k)$-close to the solutions of the continuous system, and also proving that these approximations exhibit a certain regularity. The key part here is to express the error in the potential in terms of the error in the temperature, and then only working with the temperature equation. With the given regularity, the time-discrete and fully discrete approximations can then be compared and shown to be $\mathrm{O}(h^2)$-close. The main problem here is the nonlinear term $\sigma(\theta)|\nabla\phi|^2$, which is handled in a two-step fashion: first using that $\|\nabla(\Phi_h^n - \Phi^n)\| \le C(h + \|\Theta_h^n - \Theta^n\|)$ to show that in fact $\|\nabla(\Phi_h^n - \Phi^n)\| \le Ch$ and then using this to estimate $\nabla(\Phi_h^n - \Phi^n)$ in a stronger norm.

In our case, the temperature equation (1.1) contains the extra term $\mathbf{M} : \epsilon(\dot{u})$, so our idea is to also bound the error in $\dot{u}$ by the error in the temperature. Then we show that the approximations U^n possess certain regularity, which may be used to also express the fully discrete deformation errors in terms of the fully discrete temperature errors. The key part in the latter step is to utilize the concept of Ritz-Volterra projections [22], which we here generalize to the vector-valued viscoelasticity case, as well as to discrete time.

Before we perform this extended analysis, we state the general assumptions on the given data. In these, as well as throughout the rest of the paper, C denotes a generic constant independent of k, h and n, that may differ from line to line.

Assumption 3.1. *The viscosity and elasticity tensors* $\mathbf{A}$ *and* $\mathbf{B}$ *are symmetric, and both yield Lipschitz continuous and strongly coercive bilinear forms. That is, there are positive constants* C_1, C_2 *such that for all* $u, v \in \boldsymbol{V}$ *we have*

$$\max\Big((\mathbf{A}\epsilon(u), \epsilon(v))_Q\,, (\mathbf{B}\epsilon(u), \epsilon(v))_Q \Big) \leq C_1 \|u\|_{\boldsymbol{V}} \|v\|_{\boldsymbol{V}} \quad \textit{and}$$
$$\min\Big((\mathbf{A}\epsilon(u), \epsilon(u))_Q\,, (\mathbf{B}\epsilon(u), \epsilon(u))_Q \Big) \geq C_2 \|u\|^2_{\boldsymbol{V}}.$$

Assumption 3.2. *The electrical conductivity* σ *belongs to* $C^1(\mathbb{R})$ *and there are positive constants* σ_{min}*,* σ_{max} *and* σ'_{max} *such that for all* $\theta \geq 0$ *we have*

$$0 < \sigma_{min} \leq \sigma(\theta) \leq \sigma_{max} \quad \textit{and} \quad |\sigma'(\theta)| \leq \sigma'_{max}.$$

Assumption 3.3. *The function* $f \in C(0, T;\, L^2)$*,* $\theta_0 \in H^2 \cap H^1_0$ *and* ϕ_b *is regular enough that*

$$\|\phi_b\|_{L^\infty(0,\,T;\,W^{2,12/5})} + \|\dot{\phi}_b\|_{L^2(0,\,T;\,H^1)} + \|\nabla\phi_b\|_{L^\infty(0,\,T;\,L^\infty)} \leq C.$$

By [19], these assumptions guarantee the existence of a weak solution to the problem, i.e functions (θ, ϕ, u) satisfying (2.1)–(2.3) with the time derivatives interpreted in a weak sense. Thus for example $\theta \in L^2(0, T; V)$ and $\dot{\theta} \in L^2(0, T; V)'$. For optimal convergence orders more regularity is required, and explicit conditions on the data that guarantees such regularity is currently unknown. We therefore also make the following regularity assumption:

Assumption 3.4. *There exist solutions* (θ, ϕ, u) *to* (2.1)–(2.3) *over the time interval* $[0, T]$ *which are regular enough that*

$$\|\theta\|_{L^\infty(0,\,T;\,H^2)} + \|\dot{\theta}\|_{L^\infty(0,\,T;\,L^2)} + \|\dot{\theta}\|_{L^2(0,\,T;\,H^2)} + \|\ddot{\theta}\|_{L^2(0,\,T;\,L^2)} \leq C,$$
$$\|\phi\|_{L^\infty(0,\,T;\,W^{2,12/5})} + \|\dot{\phi}\|_{L^2(0,\,T;\,H^1)} + \|\nabla\phi\|_{L^\infty(0,\,T;\,L^\infty)} \leq C,$$
$$\|\dot{u}\|_{L^\infty(0,\,T;\,H^2)} + \|\ddot{u}\|_{L^\infty(0,\,T;\,H^2)} + \|u^{(3)}\|_{L^2(0,\,T;\,L^2)} \leq C$$

The assumptions on θ and ϕ are essentially the same as in the non-deformable situation given in [21], while the assumptions on u and f are new. We note that for the non-deformable case, the existence of solutions with similar regularity properties was e.g. shown in [10] when $d \leq 2$ but with weak requirements on the initial values. Additionally, our numerical experiments suggest that in practice Assumption 3.4 is satisfied for convex domains and smooth data.

The following main theorem will be proved in the next two subsections:

Theorem 3.1. *Let Assumptions 3.1-3.4 be satisfied and let* (θ, ϕ, u) *and* $(\Theta^n_h, \Phi^n_h, U^n_h)$ *be solutions to the equations* (2.1)–(2.3) *and* (2.10)–(2.12)*, respectively. Then there are positive constants* k_0 *and* h_0 *such that if* $k < k_0$ *and* $h < h_0$ *we have for* $n = 1, \ldots, N$ *that*

$$\|\Theta^n_h - \theta_n\| + \|\Phi^n_h - \phi_n\| + \|\mathrm{D_t}\, U^n_h - \dot{u}_n\| \leq C(h^2 + k),$$

and

$$\|\Theta_h^n - \theta_n\|_{H^1} + \|\Phi_h^n - \phi_n\|_{H^1} + \|\mathrm{D_t}\, U_h^n - \dot{u}_n\|_{\boldsymbol{V}} \le C(h+k).$$

To abbreviate expressions like the above in the following, we introduce

$$e_\theta^n = \Theta^n - \theta_n, \quad e_\phi^n = \Phi^n - \phi_n \quad \text{and} \quad e_u^n = U^n - u_n$$

as well as

$$e_{\theta,h}^n = \Theta_h^n - \Theta^n, \quad e_{\phi,h}^n = \Phi_h^n - \Phi^n \quad \text{and} \quad e_{u,h}^n = U_h^n - U^n.$$

3.1. The time-discrete case.

We start by considering the semi-discrete case, and first provide a bound for $\mathrm{D_t}\, e_u^n$ in terms of e_θ^n.

Lemma 3.1. *Let Assumptions 3.1-3.4 be satisfied and let (θ, ϕ, u) and (Θ^n, Φ^n, U^n) be solutions to the equations (2.1)–(2.3) and (2.7)–(2.9), respectively. Then we have*

$$\|\mathrm{D_t}\, e_u^n\|^2 + \|e_u^n\|_{\boldsymbol{V}}^2 + k\sum_{j=1}^{n} \|\mathrm{D_t}\, e_u^j\|_{\boldsymbol{V}}^2 \le Ck^2 + Ck\sum_{j=1}^{n} \|e_\theta^j\|^2,$$

for $n = 1, \ldots, N$.

Proof. By equations (2.3) and (2.9), we see that the error e_u^n satisfies

$$\begin{aligned}\left(\mathrm{D_t^2}\, e_u^n, \boldsymbol{\chi}\right) + (\mathbf{A}\epsilon(\mathrm{D_t}\, e_u^n) + \mathbf{B}\epsilon(e_u^n), \epsilon(\boldsymbol{\chi})) &= (\mathbf{M}e_\theta^n, \epsilon(\boldsymbol{\chi})) + \left(\ddot{u}(t_n) - \mathrm{D_t^2}\, u(t_n), \boldsymbol{\chi}\right) \\ &\quad + (\mathbf{A}\epsilon(\dot{u}(t_n) - \mathrm{D_t}\, u(t_n)), \epsilon(\boldsymbol{\chi})) \\ &\le C\|e_\theta^n\|\|\boldsymbol{\chi}\|_{\boldsymbol{V}} + Ck\|\boldsymbol{\chi}\| + Ck\|\boldsymbol{\chi}\|_{\boldsymbol{V}}\end{aligned}$$

due to the regularity assumptions on u. We note that for any sequence $\{g^n\}$ we have

$$2\left({\mathrm{D_t}}^2 g^n, \mathrm{D_t}\, g^n\right) \ge \mathrm{D_t}\|\mathrm{D_t}\, g^n\|^2 \quad \text{and} \quad 2\left(\mathbf{B}\epsilon(g^n), \mathrm{D_t}\, g^n\right) \ge \mathrm{D_t}\|\mathrm{D_t}\, g^n\|_{\mathbf{B}}^2,$$

where $\|\cdot\|_{\mathbf{B}}$ is the norm induced by the inner product $(\mathbf{B}\epsilon(\cdot), \epsilon(\cdot))$. Thus by choosing $\boldsymbol{\chi} = \mathrm{D_t}\, e_u^n$ and using the Cauchy–Schwarz inequality as well as Young's inequality, $ab \le \frac{1}{2c}a^2 + \frac{c}{2}b^2$, we get

$$\mathrm{D_t}\|\mathrm{D_t}\, e_u^n\|^2 + 2C_2\|\mathrm{D_t}\, e_u^n\|_{\boldsymbol{V}} + \mathrm{D_t}\|e_u^n\|_{\mathbf{B}}^2 \le Ck^2 + C\|e_\theta^n\|^2 + C_2\|\mathrm{D_t}\, e_u^n\|_{\boldsymbol{V}}^2.$$

Canceling the final term, summing over n and modifying the constants then yields

$$\|\mathrm{D_t}\, e_u^n\|^2 + k\sum_{j=1}^{n} \|\mathrm{D_t}\, e_u^j\|_{\boldsymbol{V}} + \|e_u^n\|_{\mathbf{B}}^2 \le Ck^2 + Ck\sum_{j=1}^{n} \|e_\theta^j\|^2,$$

and the Lemma follows from the equivalence between the $\mathbf{B}$- and $\boldsymbol{V}$-norms.

□ □

Theorem 3.2. *Let Assumptions 3.1-3.4 be satisfied and let (θ, ϕ, u) and (Θ^n, Φ^n, U^n) be solutions to the equations (1.1)–(1.3) and (2.4)–(2.6), respectively. Then there is a positive constant k_0 such that if $k < k_0$ then*

$$\|e_\theta^n\|_{H^1}^2 + \|e_\phi^n\|_{H^1}^2 + \|\mathrm{D_t}\, e_u^n\|_{\boldsymbol{V}}^2 \le Ck^2,$$

for $n = 1, \dots, N$. In addition, the approximations have the following regularity:

$$\|\Theta^n\|_{H^2}^2 + \|\mathrm{D_t}\,\Theta^n\|^2 + k\sum_{j=1}^{n}\|\mathrm{D_t}\,\Theta^j\|_{H^2}^2 \le C,$$
$$\|\Phi^n\|_{W^{2,12/5}} + \|\nabla\Phi^n\|_{L^\infty} \le C,$$
$$\|\mathrm{D_t}\,U^n\|_{H^2}^2 + \|\mathrm{D_t^2}\,U^n\|_{\boldsymbol{V}}^2 + k\sum_{j=1}^{n}\|\mathrm{D_t^2}\,U^j\|_{H^2}^2 \le C.$$

Proof. To begin with, we see that the error e_ϕ^n satisfies

$$-\nabla\cdot\big(\sigma(\Theta^n)\nabla e_\phi^n\big) = \nabla\cdot\big((\sigma(\Theta^n) - \sigma(\theta_n))\nabla\phi_n\big).$$

Multiplying this equation by e_ϕ^n and integrating directly yields

$$\|\nabla e_\phi^n\|^2 \le C\|\nabla\phi_n\|_{L^\infty}\|e_\theta^n\|\|\nabla e_\phi^n\|,$$

so that

$$\|\nabla e_\phi^n\| \le C\|e_\theta^n\| \tag{3.1}$$

by the regularity assumptions. This inequality for e_ϕ^n corresponds to Lemma 3.1 for e_u^n. Further, we see that the error e_θ^n satisfies

$$\begin{aligned}\mathrm{D_t}\,e_\theta^n - \Delta e_\theta^n = {} & \Big(\sigma(\Theta^{n-1}) - \sigma(\theta_{n-1})\Big)|\nabla\phi_{n-1}|^2 + \sigma(\Theta^{n-1})\Big(\nabla\Phi^{n-1} + \nabla\phi_{n-1}\Big)\cdot\nabla e_\phi^{n-1} \\ & - M : \epsilon(\mathrm{D_t}\,e_u^{n-1}) + R_\theta^n,\end{aligned} \tag{3.2}$$

where

$$\begin{aligned}R_\theta^n = {} & \big(\sigma(\theta_{n-1}) - \sigma(\theta_n)\big)|\nabla\phi_{n-1}|^2 + \sigma(\theta_n)\big(\nabla\phi_{n-1} + \nabla\phi_n\big)\cdot\big(\nabla\phi_{n-1} - \nabla\phi_n\big) \\ & + M : \epsilon(\dot{u}_n - \dot{u}_{n-1}) + M : \epsilon(\dot{u}_{n-1} - \mathrm{D_t}\,u_{n-1}).\end{aligned}$$

is bounded by $\|R_\theta^n\| \le Ck$, again by the regularity assumptions. By multiplying by e_θ^n and integrating, we therefore get

$$\begin{aligned}\mathrm{D_t}\|e_\theta^n\|^2 + 2\|\nabla e_\theta^n\|^2 \le {} & C\|e_\theta^{n-1}\|\|e_\theta^n\|\|\nabla\phi_{n-1}\|_{L^\infty} + \big(M : \epsilon(\mathrm{D_t}\,e_u^{n-1}), e_\theta^n\big) + Ck\|e_\theta^n\| \\ & + \Big(\sigma(\Theta^{n-1})\big(\nabla\Phi^{n-1} + \nabla\phi_{n-1}\big)e_\theta^n, \nabla e_\phi^{n-1}\Big).\end{aligned} \tag{3.3}$$

The last term of this expression can be shown to be bounded by $C(\|e_\theta^n\|^2 + \|e_\phi\|_{H^1}^2)$, see [21, p.627], and for the second we observe that for a generic $u \in \boldsymbol{V}$,

$$(\mathbf{M} : (\nabla u), \chi)_{L^2} = (\nabla u, \mathbf{M}\chi)_Q = -\,(u, \nabla\cdot(\mathbf{M}\chi))_{\boldsymbol{L}^2} = -\,(u, \mathbf{M}\nabla\chi)_{\boldsymbol{L}^2}.$$

As a completely analogous calculation holds also for $(\nabla u)^T$ and $\mathbf{M}$ is symmetric, we thus have

$$(\mathbf{M} : \epsilon(u), \chi) = -\,(u, \mathbf{M}\nabla\chi) \le C\|u\|\|\nabla\chi\|. \tag{3.4}$$

This implies that (3.3) reduces to

$$\mathrm{D_t}\|e_\theta^n\|^2 + 2\|\nabla e_\theta^n\|^2 \le C\big(k^2 + \|e_\theta^{n-1}\|^2 + \|e_\theta^n\|^2 + \|e_\phi^{n-1}\|_{H^1}^2 + \|\mathrm{D_t}\,e_u^{n-1}\|^2\big) + \|\nabla e_\theta^n\|^2.$$

Canceling the last term, summing up and using Equation (3.1) and Lemma 3.1 thus yields

$$\|e_\theta^n\|^2 + k\sum_{j=1}^{n}\|\nabla e_\theta^j\|^2 \le Ck^2 + Ck\sum_{j=1}^{n}\|e_\theta^j\|^2.$$

Under the step size restriction $Ck < 1$, we can eliminate the last term of the sum. An application of Grönwall's lemma then shows that the left-hand side is bounded by Ck^2. Using Equation (3.1) and Lemma 3.1 again, we see that in fact

$$\|e_\theta^n\|^2 + k\sum_{j=1}^{n} \|\nabla e_\theta^j\|^2 + \|\nabla e_\phi^n\|^2 + \|\mathrm{D_t}\, e_u^n\|^2 + \|e_u^n\|_{\boldsymbol{V}}^2 + k\sum_{j=1}^{n} \|\mathrm{D_t}\, e_u^j\|_{\boldsymbol{V}}^2 \le Ck^2$$

From these preliminary bounds, we may deduce the desired regularity of Θ^n and Φ^n and then test (3.2) with $-\Delta e_\theta^n$ to acquire

$$\|e_\theta^n\|_{H^1}^2 + k\sum_{j=1}^{n} \|\Delta e_\theta^j\|^2 \le Ck^2.$$

For details, we refer to [21, Theorem 3.1]. Let us instead investigate the remaining questions of the regularity of U^n and the pointwise bound for $\mathrm{D_t}\, e_u^n$ in the $\boldsymbol{V}$-norm. By the defining equation, we have that

$$\begin{aligned} \nabla\cdot\big(\mathbf{A}\epsilon(\mathrm{D_t}\, e_u^n) + \mathbf{B}\epsilon(e_u^n)\big) = \mathrm{D_t^2}\, e_u^n + \nabla\cdot\big(\mathbf{M}\Theta^n\big) + \mathrm{D_t}^2\, u(t_n) - \ddot{u}(t_n) \\ + \nabla\cdot\big(\mathbf{A}\epsilon(\mathrm{D_t}\, u(t_n) - \dot{u}(t_n))\big), \end{aligned} \tag{3.5}$$

where the right-hand side is in L^2 since $\|\mathrm{D_t^2}\, e_u^n\| \le k^{-1}(\|\mathrm{D_t}\, e_u^n\| + \|\mathrm{D_t}\, e_u^{n-1}\|) \le C$. Let us denote it by g_n. Then we can rewrite the previous equation as

$$\nabla\cdot\big(\mathbf{A}\epsilon(\mathrm{D_t}\, e_u^n) + k\mathbf{B}\epsilon(\mathrm{D_t}\, e_u^n)\big) = g_n + \nabla\cdot\big(\mathbf{B}\epsilon(e_u^{n-1})\big).$$

Now since both $\mathbf{B}$ and $\mathbf{A}+k\mathbf{B}$ induces bounded and coercive inner products on $\boldsymbol{V}$, we see that

$$\begin{aligned} \|\mathrm{D_t}\, e_u^n\|_{H^2}^2 &\le C\|\nabla\cdot\big(\mathbf{A}\epsilon(\mathrm{D_t}\, e_u^n) + k\mathbf{B}\epsilon(\mathrm{D_t}\, e_u^n)\big)\|^2 \\ &\le C\|g_n\|^2 + C\|e_u^{n-1}\|_{H^2}^2 \end{aligned}$$

But since $e_u^{n-1} = k\sum_{j=1}^{n-1} \mathrm{D_t}\, e_u^j$, we can estimate the second term by Cauchy–Schwarz as

$$\|e_u^{n-1}\|_{H^2}^2 \le k\sum_{j=1}^{n-1} \|\mathrm{D_t}\, e_u^j\|_{H^2}^2.$$

An application of Grönwall's lemma thus shows that

$$\|\mathrm{D_t}\, e_u^n\|_{H^2} \le C,$$

which also implies that e_u^n, U^n and $\mathrm{D_t}\, U^n$ are all in H^2. We may now multiply (3.5) by $\nabla\cdot\big((\mathbf{A}+k\mathbf{B})\epsilon(\mathrm{D_t}\, e_u^n)\big)$ and integrate to get

$$(\mathrm{D_t}\, \epsilon(\mathrm{D_t}\, e_u^n), (\mathbf{A}+k\mathbf{B})\epsilon(\mathrm{D_t}\, e_u^n)) + \|\nabla\cdot\big((\mathbf{A}+k\mathbf{B})\epsilon(\mathrm{D_t}\, e_u^n)\big)\|^2 \le C\|e_\theta^n\|_{H^1}^2 + C\|e_\theta^{n-1}\|_{H^2}^2,$$

where we have used the Cauchy-Schwarz and Young inequalities and canceled a term $\frac{1}{2}\|\nabla\cdot\big((\mathbf{A}+k\mathbf{B})\epsilon(\mathrm{D_t}\, e_u^n)\big)\|^2$. The first term on the left-hand side can be estimated from below by $\mathrm{D_t}\|\mathrm{D_t}\, e_u^n\|_{\mathbf{A}+k\mathbf{B}}$, so summing up and using the equivalence of the $(\mathbf{A}+k\mathbf{B})$- and $\boldsymbol{V}$-norms, we get

$$\|\mathrm{D_t}\, e_u^n\|_{\boldsymbol{V}}^2 + k\sum_{j=1}^{n} \|\mathrm{D_t}\, e_u^j\|_{H^2}^2 \le Ck\sum_{j=1}^{n-1} \|e_\theta^j\|_{H^1}^2 + Ck\sum_{j=1}^{n-1} \|\mathrm{D_t}\, e_u^j\|_{H^2}^2.$$

But the first term in the right-hand side is bounded by Ck^2 and in the second we may again use that $\|\mathrm{D}_{\mathrm{t}}\, e_u^j\|_{H^2}^2 \leq k\sum_{i=1}^{j} \|\mathrm{D}_{\mathrm{t}}\, e_u^i\|_{H^2}^2$. Defining

$$w_n = \|\mathrm{D}_{\mathrm{t}}\, e_u^n\|_{\boldsymbol{V}}^2 + k\sum_{j=1}^{n} \|\mathrm{D}_{\mathrm{t}}\, e_u^j\|_{H^2}^2,$$

we thus have

$$w_n \leq Ck^2 + Ck\sum_{j=1}^{n-1} w_j,$$

and an application of Grönwall's lemma shows that $w_n \leq Ck^2$. This yields the final desired error bound, and additionally shows that $\|{\mathrm{D}_{\mathrm{t}}}^2\, e_u^n\|_{\boldsymbol{V}}^2 + k\sum_{j=1}^{n} \|{\mathrm{D}_{\mathrm{t}}}^2\, e_u^j\|_{H^2}^2 \leq C$ which implies the stated regularity for U^n.

□ □

3.2. **The fully discrete case.** We now turn to the fully discretized case and first prove an analogue to Lemma 3.1.

Lemma 3.2. *Let Assumptions 3.1-3.4 be satisfied and* (Θ^n, Φ^n, U^n) *and* $(\Theta_h^n, \Phi_h^n, U_h^n)$ *be solutions to equations* (2.7)–(2.9) *and* (2.10)–(2.12)*, respectively. Then there is a positive constant* k_0 *such that if* $k < k_0$ *we have for* $n = 1, \ldots, N$ *that*

$$\|e_{u,h}^n\|^2 + \|\mathrm{D}_{\mathrm{t}}\, e_{u,h}^n\|^2 \leq Ch^4 + Ck\sum_{j=1}^{n} \|e_{\theta,h}^j\|^2 \quad \textit{and}$$

$$\|e_{u,h}^n\|_{\boldsymbol{V}}^2 + k\sum_{j=1}^{n} \|\mathrm{D}_{\mathrm{t}}\, e_{u,h}^j\|_{\boldsymbol{V}}^2 \leq Ch^2 + Ck\sum_{j=1}^{n} \|e_{\theta,h}^j\|^2.$$

Remark 3.1. In the case of a first-order equation, one would typically first add and subtract the Ritz projection of e_u^n in order to work only in the finite element space. This approach is viable also in the second-order case, if one defines the Ritz projection using the $(\mathbf{A}\epsilon(\cdot), \epsilon(\cdot))$ inner product. We refer to [25] for the scalar-valued case. However, we choose to instead work with a Ritz-Volterra projection, see [22] for the scalar-valued case. Such a projection takes both the $\mathbf{A}$- and $\mathbf{B}$-terms into account simultaneously, i.e. it is a projection of $C^1(0,T;\boldsymbol{V})$-functions rather than of elements in $\boldsymbol{V}$. In the present situation, we need of course to consider a discretized version, but it nevertheless simplifies matters.

Proof. Subtracting (2.9) from (2.12), we see that

$$\big(\mathrm{D}_{\mathrm{t}}^2\, e_{u,h}^n, \boldsymbol{\chi}\big) + \big(\mathbf{A}\epsilon(\mathrm{D}_{\mathrm{t}}\, e_{u,h}^n) + \mathbf{B}\epsilon(e_{u,h}^n), \epsilon(\boldsymbol{\chi})\big) = \big(\mathbf{M}e_{\theta,h}^n, \epsilon(\boldsymbol{\chi})\big)$$

for all $\boldsymbol{\chi} \in \boldsymbol{S}_h$. Now let $e_{u,h}^n = \eta^n + \rho^n$, where

$$\eta^n = U_h^n - W^n \in \boldsymbol{S}_h \quad \text{and} \quad \rho^n = W^n - U^n,$$

with the discrete Ritz-Volterra projection W^n of U^n satisfying $W^0 = U^0 = 0$ and

$$(\mathbf{A}\epsilon(\mathrm{D}_{\mathrm{t}}\, W^n - \mathrm{D}_{\mathrm{t}}\, U^n) + \mathbf{B}\epsilon(W^n - U^n), \epsilon(\boldsymbol{\chi})) = 0 \tag{3.6}$$

for all $\boldsymbol{\chi} \in \boldsymbol{S}_h$. We note that Equation (3.6) may also be stated as

$$(\mathbf{A}\epsilon(\mathrm{D}_{\mathrm{t}}\, \rho^n) + \mathbf{B}\epsilon(\rho^n), \epsilon(\boldsymbol{\chi})) = 0,$$

and that since $U^1 = 0$, also $W^1 = 0$. Additionally, we need the Ritz projection R_h given by the viscosity term. For a generic $u \in \boldsymbol{V}$, this is defined by

$$(\mathbf{A}\epsilon(R_h u - u), \epsilon(\boldsymbol{\chi})) = 0$$

for all $\boldsymbol{\chi} \in \boldsymbol{S}_h$, and we have the inequality

$$\|R_h u - u\| + h\|R_h u - u\|_{\boldsymbol{V}} \leq Ch^2\|u\|_{H^2}.$$

We start by estimating the $\boldsymbol{V}$-norms of $\mathrm{D_t}\,\rho^n$ and ρ^n. To this end, we observe that for a generic u, we have

$$\|u\|_{\boldsymbol{V}}^2 = \|\epsilon(u)\|_Q^2 \leq \|\nabla u\|_Q^2 = \sum_{j=1}^{d} \Big\|\frac{\partial u}{\partial x_j}\Big\|^2$$

and that

$$\Big\|\frac{\partial u}{\partial x_j}\Big\| = \sup_{\varphi \in C_0^\infty(\Omega)^d, \|\varphi\|=1} \left(\frac{\partial u}{\partial x_j}, \varphi\right).$$

We therefore take $\varphi \in C_0^\infty(\Omega)^d$ with $\|\varphi\| = 1$ and let $\Psi \in \boldsymbol{V}$ be the solution to

$$(\mathbf{A}\epsilon(\Psi), \epsilon(\boldsymbol{\chi}))_Q = -\left(\frac{\partial \varphi}{\partial x_j}, \boldsymbol{\chi}\right).$$

Then

$$\begin{aligned}
\left(\frac{\partial\,\mathrm{D_t}\,\rho^n}{\partial x_j}, \varphi\right) &= -\left(\mathrm{D_t}\,\rho^n, \frac{\partial \varphi}{\partial x_j}\right) = (\mathbf{A}\epsilon(\Psi), \epsilon(\mathrm{D_t}\,\rho^n)) = (\mathbf{A}\epsilon(\mathrm{D_t}\,\rho^n), \epsilon(\Psi)) \\
&= (\mathbf{A}\epsilon(\mathrm{D_t}\,\rho^n), \epsilon(\Psi - R_h\Psi)) + (\mathbf{A}\epsilon(\mathrm{D_t}\,\rho^n), \epsilon(R_h\Psi)) \\
&= (\mathbf{A}\epsilon(\mathrm{D_t}\,\rho^n), \epsilon(\Psi - R_h\Psi)) - (\mathbf{B}\epsilon(\rho^n), \epsilon(R_h\Psi)) =: R_1 + R_2,
\end{aligned}$$

where the last term is bounded by

$$R_2 \leq C\|\rho^n\|_{\boldsymbol{V}}\|R_h\Psi\|_{\boldsymbol{V}} \leq C\|\rho^n\|_{\boldsymbol{V}}(\|R_h\Psi - \Psi\|_{\boldsymbol{V}} + \|\Psi\|_{\boldsymbol{V}}) \leq C\|\rho^n\|_{\boldsymbol{V}}.$$

Moreover, since $\mathrm{D_t}\,W^n \in \boldsymbol{S}_h$, the first term is bounded by

$$\begin{aligned}
R_1 = -\,(\mathbf{A}\epsilon(\mathrm{D_t}\,U^n), \epsilon(\Psi - R_h\Psi)) &= (\mathbf{A}\epsilon(R_h\,\mathrm{D_t}\,U^n - \mathrm{D_t}\,U^n), \epsilon(\Psi - R_h\Psi)) \\
&= (\mathbf{A}\epsilon(R_h\,\mathrm{D_t}\,U^n - \mathrm{D_t}\,U^n), \epsilon(\Psi)) \\
&\leq C\|R_h\,\mathrm{D_t}\,U^n - \mathrm{D_t}\,U^n\|_{\boldsymbol{V}}\|\Psi\|_{\boldsymbol{V}} \\
&\leq Ch\|\mathrm{D_t}\,U^n\|_{H^2}.
\end{aligned}$$

By expressing ρ^n in terms of $\mathrm{D_t}\,\rho^j$ and noting that $\rho^0 = 0$, we thus have

$$\|\mathrm{D_t}\,\rho^n\|_{\boldsymbol{V}} \leq Ch\|\mathrm{D_t}\,U^n\|_{H^2} + Ck\sum_{j=1}^{n}\|\mathrm{D_t}\,\rho^j\|_{\boldsymbol{V}},$$

and under the step size restriction $Ck < 1$ we can eliminate the last term of the sum and apply Grönwall's lemma. This shows that

$$\|\mathrm{D_t}\,\rho^n\|_{\boldsymbol{V}} \leq Ch\Big(\|\mathrm{D_t}\,U^n\|_{H^2} + Ck\sum_{j=1}^{n-1}\|\mathrm{D_t}\,U^j\|_{H^2}\Big).$$

By using the regularity shown in Theorem 3.2 and then summing over n, we see that

$$\|\rho^n\|_{\boldsymbol{V}} + \|\mathrm{D_t}\,\rho^n\|_{\boldsymbol{V}} \leq Ch.$$

Using these bounds we may now estimate ρ also in the L^2-norm, by instead letting $\Psi \in \boldsymbol{V}$ be the solution to

$$(\mathbf{A}\epsilon(\Psi), \epsilon(\boldsymbol{\chi}))_Q = -\,(\varphi, \boldsymbol{\chi}).$$

Then as before,

$$(\mathrm{D_t}\,\rho^n, \varphi) = (\mathbf{A}\epsilon(R_h\,\mathrm{D_t}\,U^n - \mathrm{D_t}\,U^n), \epsilon(\Psi)) + (\mathbf{B}\epsilon(\rho^n), \epsilon(R_h\Psi)) =: R_3 + R_4,$$

where

$$R_3 \le C\|R_h\,\mathrm{D_t}\,U^n - \mathrm{D_t}\,U^n\|_{\boldsymbol{V}}\|\Psi\|_{\boldsymbol{V}} \le Ch^2\|\mathrm{D_t}\,U^n\|_{H^2}.$$

For R_4, we note that $\|\Psi\|_{H^2} \le C\|\varphi\| \le C$, so that by using integration by parts and observing that both ρ^n and Ψ are zero on $\partial\Omega$ we get,

$$\begin{aligned} R_4 &\le (\mathbf{B}\epsilon(\rho^n), \epsilon(R_h\Psi - \Psi)) + (\mathbf{B}\epsilon(\rho^n), \epsilon(\Psi)) \\ &\le C\|\rho^n\|_{\boldsymbol{V}}\|R_h\Psi - \Psi\|_{\boldsymbol{V}} + C\|\rho^n\|\|\Psi\|_{H^2} + \|\rho^n\|_{L^2(\partial\Omega)}\|\Psi\|_{H^1(\partial\Omega)} \\ &\le Ch^2 + C\|\rho^n\|. \end{aligned}$$

Hence similarly to the calculation for the $\boldsymbol{V}$-norm, Grönwall's lemma implies that

$$\|\mathrm{D_t}\,\rho^n\| \le Ch^2\Big(\|\mathrm{D_t}\,U^n\|_{H^2} + Ck\sum_{j=1}^{n-1}\|\mathrm{D_t}\,U^j\|_{H^2}\Big),$$

so that

$$\|\rho^n\| + \|\mathrm{D_t}\,\rho^n\| \le Ch^2.$$

To bound η^n, we also need a bound on the second derivative of ρ^n. For this, we apply $\mathrm{D_t}$ to (3.6) and then follow the same procedure as above. This shows that

$$\|\mathrm{D_t^2}\,\rho^n\|_{\boldsymbol{V}} \le Ch\Big(\|\mathrm{D_t^2}\,U^n\|_{H^2} + Ck\sum_{j=1}^{n-1}\|\mathrm{D_t^2}\,U^j\|_{H^2}\Big),$$

and similarly for the L^2-norm, but with h^2 instead of h. We do not have pointwise H^2-regularity of $\mathrm{D_t^2}\,U^n$ from Theorem 3.2, but we may estimate the sum by

$$k\sum_{j=1}^{n-1}\|\mathrm{D_t^2}\,U^j\|_{H^2} \le \Big(k\sum_{j=1}^{n-1}\|\mathrm{D_t^2}\,U^j\|_{H^2}^2\Big)^{1/2} \le C,$$

and conclude that

$$\|\mathrm{D_t^2}\,\rho^n\| + h\|\mathrm{D_t^2}\,\rho^n\|_{\boldsymbol{V}} \le Ch^2 + Ch^2\|\mathrm{D_t^2}\,U^n\|_{H^2}. \tag{3.7}$$

Here the $\|\mathrm{D_t^2}\,U^n\|_{H^2}$-term is not necessarily finite, but since this bound will only be used inside a sum it causes no problems.

Now for η^n, by using (3.6) to exchange W^n for U^n and then (2.9), (2.12), we get

$$\begin{aligned} &\big(\mathrm{D_t^2}\,\eta^n, \boldsymbol{\chi}\big) + (\mathbf{A}\epsilon(\mathrm{D_t}\,\eta^n) + \mathbf{B}\epsilon(\eta^n), \epsilon(\boldsymbol{\chi})) \\ &\qquad = \big(\mathrm{D_t^2}\,U^n - \mathrm{D_t^2}\,W^n, \boldsymbol{\chi}\big) + \big(\mathbf{M}e_{\theta,h}^n, \epsilon(\boldsymbol{\chi})\big) \\ &\qquad = -\big(\mathrm{D_t^2}\,\rho^n, \boldsymbol{\chi}\big) + \big(Me_{\theta,h}^n, \epsilon(\boldsymbol{\chi})\big) \end{aligned}$$

Choosing $\boldsymbol{\chi} = \mathrm{D_t}\,\eta^n \in \boldsymbol{S}_h$, by (3.7) we get, after canceling a $C_2\|\mathrm{D_t}\,\eta^n\|_{\boldsymbol{V}}^2$ term,

$$\mathrm{D_t}\|\mathrm{D_t}\,\eta^n\|^2 + C_2\|\mathrm{D_t}\,\eta^n\|_{\boldsymbol{V}}^2 + \mathrm{D_t}\|\eta^n\|_{\mathbf{B}}^2 \le C\big(h^4 + h^4\|\mathrm{D_t^2}\,U^n\|_{H^2}^2 + \|e_{\theta,h}^n\|^2\big),$$

so summing and noting again that $k\sum_{j=1}^{n-1}\|\mathrm{D_t}\,U^j\|_{H^2}^2 \le C$, we have

$$\|\mathrm{D_t}\,\eta^n\|^2 + k\sum_{j=1}^{n-1}\|\mathrm{D_t}\,\eta^j\|_{\boldsymbol{V}}^2 + \|\eta^n\|_{\boldsymbol{V}}^2 \le Ch^4 + Ck\sum_{j=1}^{n-1}\|e_{\theta,h}^j\|^2.$$

Finally, combining the bounds for ρ^n, η^n and their first derivatives leads to the statement of the lemma. □ □

Remark 3.2. We note that the regularity given in Theorem 3.2 is not enough to show $\|\mathrm{D_t}\, e_{u,h}^n\|_{\boldsymbol{V}}^2 \leq Ch^2 + Ck\sum_{j=1}^n \|e_{\theta,h}^j\|^2$, but such a bound is not required for the proof of the next theorem.

Theorem 3.3. *Let Assumptions 3.1-3.4 be satisfied and* (Θ^n, Φ^n, U^n) *and* $(\Theta_h^n, \Phi_h^n, U_h^n)$ *be solutions to equations* (2.7)–(2.9) *and* (2.10)–(2.12)*, respectively. Then there are positive constants* k_0 *and* h_0 *such that if* $k < k_0$ *and* $h < h_0$ *then for* $n = 1, \dots, N$,

$$\|e_{\theta,h}^n\| + \|e_{\phi,h}^n\| + \|\mathrm{D_t}\, e_{u,h}^n\| \leq Ch^2 \quad \textit{and} \quad \|e_{\theta,h}^n\|_{H^1} + \|e_{\phi,h}^n\|_{H^1} + \|\mathrm{D_t}\, e_{u,h}^n\|_{\boldsymbol{V}} \leq Ch.$$

Proof. The idea is, similarly to the time-discrete case, essentially to write down the equation for e_θ^n, test it with e_θ^n, express the errors $e_{u,h}$ and $e_{\phi,h}$ in terms of $e_{\theta,h}$ by Lemma 3.2 and its potential-analogue, and finally use Grönwall's lemma. However, since e_θ^n does not belong to the finite element space, we need to introduce instead

$$e_h^n = \Theta_h^n - R_h\Theta^n,$$

where $\|e_{\theta,h}^n\| \leq \|e_h^n\| + \|R_h\Theta^n - \Theta^n\| \leq \|e_h^n\| + Ch^2$ due to Theorem 3.2. With this definition, we see that for all $\chi \in S_h$,

$$(\mathrm{D_t}\, e_h^n, \chi) + (\nabla\theta_h^n, \nabla\chi) = (\mathrm{D_t}(\Theta^n - R_h\Theta^n), \chi) + (R_\phi, \chi) - \Big(M : \epsilon(\mathrm{D_t}\, e_{u,h}^{n-1}), \chi\Big),$$

where R_ϕ contains terms related to the potential ϕ. Choosing $\chi = e_h^n$, we know from [21] that

$$(R_\phi, e_h^n) \leq Ch^3 + Ch^4\|\mathrm{D_t}\,\Theta^n\|_{H^2}^2 + Ch^{-1}\|e_h^{n-1}\|^4 + C\|e_h^{n-1}\|^2 + \frac{1}{4}\|e_h^n\|_{H^1}^2,$$

and we also have by (3.4) that

$$\Big(M : \epsilon(\mathrm{D_t}\, e_{u,h}^{n-1}), e_h^n\Big) \leq C\|\mathrm{D_t}\, e_{u,h}^{n-1}\|^2 + \frac{1}{4}\|e_h^n\|_{H^1}^2.$$

We additionally know that $\|e_h^0\| = \|I_h\theta_0 - \theta_0\| \leq Ch^2 < h^{1/2}$ if $h < h_0$. Assuming that $\|e_h^m\| \leq h^{1/2}$ for $m = 1, \dots, n-1$ therefore means that

$$\mathrm{D_t}\|e_h^m\|^2 + \|e_h^m\|_{H^1}^2 \leq Ch^3 + Ch^4\|\mathrm{D_t}\,\Theta^m\|_{H^2}^2 + C\|e_h^{m-1}\|^2 + C\|\mathrm{D_t}\, e_{u,h}^{m-1}\|^2$$

for $m = 1, \dots, n$, which after summation and usage of Lemma 3.2 yields

$$\begin{aligned}\|e_h^m\|^2 + k\sum_{j=1}^m \|e_h^j\|_{H^1}^2 &\leq Ch^3 + Ch^4 + Ck\sum_{j=1}^{m-1}\|e_h^j\|^2 + Ck\sum_{j=1}^{m-1}\|\mathrm{D_t}\, e_{u,h}^j\|^2 \\ &\leq Ch^3 + Ck\sum_{j=1}^{m-1}\Big(\|e_h^j\|^2 + Ck\sum_{i=1}^{j}\|e_h^i\|^2\Big).\end{aligned}$$

If we now set $g_m = \max_{1\leq j\leq m}\left(\|e_h^j\|^2 + Ck\sum_{i=1}^j \|e_h^i\|^2\right)$ we have

$$g_m \leq Ch^3 + Ck\sum_{j=1}^{m-1} g_j,$$

to which we may apply Grönwall's lemma to acquire

$$\|e_h^n\|^2 + Ck\sum_{j=1}^n \|e_h^j\|^2 \leq \tilde{C}h^3.$$

Hence if $\tilde{C}h^{5/2} \leq 1$ we have that $\|e_h^n\| \leq h^{1/2}$. Thus by induction $\|e_h^n\| \leq h^{1/2}$ holds for all n such that $0 \leq n \leq N$. But then also the other calculations just performed are valid for $1 \leq n \leq N$, so in fact $\|e_h^n\| \leq h^{3/2}$. This preliminary bound may be used as in [21, p.631] to show $\|e_{\phi,h}^n\| \leq Ch$ and to improve the bound of the quadratic potential term to

$$(R_\phi, e_h^n) \leq Ch^4 + Ch^4 \|\mathrm{D_t}\, \Theta^n\|_{H^2}^2 + C\|e_h^{n-1}\|^2 + \frac{1}{4}\|e_h^n\|_{H^1}^2.$$

Hence,

$$\|e_h^n\|^2 + k\sum_{j=1}^{n} \|e_h^j\|_{H^1}^2 \leq Ch^4 + Ck\sum_{j=1}^{m-1} \Big(\|e_h^j\|^2 + Ck\sum_{i=1}^{j} \|e_h^i\|^2\Big),$$

and once more applying Grönwall's lemma to g_n shows that

$$\|e_h^n\|^2 + k\sum_{j=1}^{n} \|e_h^j\|_{H^1}^2 \leq Ch^4.$$

This proves $\|e_{\theta,h}^n\| \leq Ch^2$, and from [21] we find $\|e_{\phi,h}^n\| + h\|e_{\phi,h}^n\|_{H^1} \leq Ch^2$. Applying Lemma 3.2 gives $\|\mathrm{D_t}\, e_{u,h}^n\| \leq Ch^2$. Finally, by inverse inequalities we find also that $\|e_{\theta,h}^n\|_{H^1} + \|\mathrm{D_t}\, e_{u,h}^n\|_{\boldsymbol{V}} \leq Ch$. □ □

of Theorem 3.1. This follows directly from Theorem 3.2 and Theorem 3.3 upon observing that, e.g.,

$$\|\mathrm{D_t}\, U_h^n - \dot{u}_n\| \leq \|e_{u,h}\| + \|e_u\| + \|\mathrm{D_t}\, u_n - \dot{u}_n\|,$$

where the last term is bounded in the proper way due to the regularity assumptions on the solution to the continuous system. □ □

4. Numerical experiments

We have implemented both the method based on (2.10)–(2.12) and the corresponding fully implicit method based on implicit Euler, using FEniCS (see e.g. [4, 23]). These implementations were then used to verify our theoretical results by applying them to the following test examples.

4.1. **Problem 1.**

First consider the two-dimensional problem with $\Omega = (0,1)^2$, $\mathbf{M} = I$, $f = [0,0]^T$ and the viscosity and elasticity tensors given in Voigt notation by

$$\mathbf{A} = \mathbf{B} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

We take the electrical conductivity to be given by

$$\sigma(\theta) = 2.5 - \arctan(5\theta - 10),$$

which has a rather steep slope close to $\theta = 2$. The initial conditions are given by $\theta_0(x,y) = 0$ and $u_0(x,y) = v_0(x,y) = [0,0]^T$. These functions also define the Dirichlet boundary conditions for θ and u, while for ϕ they are given by $\phi_b(x,y) = 5(1-x)$.

We discretize Ω by first subdividing it into squares and then dividing each square into four triangles. With N_x squares in each dimension, each triangle has diameter $h = 1/N_x$ and the full grid has $4N_x^2$ triangles. We take $N_x \in \{4, 8, 16, 32, 64\}$. Since the error should be $\mathrm{O}(h^2 + k)$, we choose the number of time steps to be $N_t = N_x^2/2$. With the final time $T = 1$, this gives $k = 2h^2$. We emphasize here that the time steps could be taken much larger than this, but illustrating the error

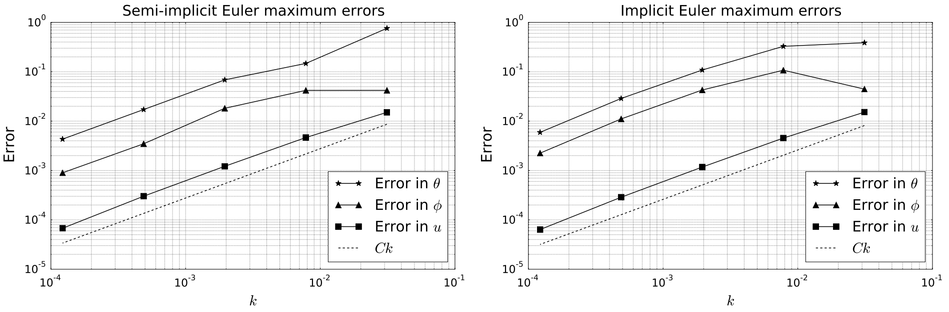

FIGURE 1. The errors (4.1) for the problem defined in Section 4.1, computed by the semi-implicit method (left) and the implicit Euler method (right).

is then less straightforward. Finally, because the exact solution of the problem is not available we cannot compute the exact errors. Instead, we compare the different approximations to a reference approximation $(\Theta_{\text{ref}}, \Phi_{\text{ref}}, U_{\text{ref}})$ computed by the implicit Euler scheme with $N_x = 128$ and $N_t = 8192$.

Figure 1 shows the errors

$$\max_{1\le n\le N_t} \|\Theta_h^n - \Theta_{\text{ref}}(t_n)\|_{L^2}, \max_{1\le n\le N_t} \|\Phi_h^n - \Phi_{\text{ref}}(t_n)\|_{L^2} \quad \text{and} \quad \max_{1\le n\le N_t} \|U_h^n - U_{\text{ref}}(t_n)\|_{L^2} \tag{4.1}$$

for the different discretizations on a logarithmic scale, for both the semi-implicit method (left) and the method based on implicit Euler (right). These clearly exhibit the expected error behaviour predicted by Theorem 3.3, except for the first points where the grid is very coarse. We also note that the errors are very similar in size, which means that the semi-implicit method is much more efficient. A peculiar effect in this case is that the semi-implicit errors in θ and ϕ are actually less than the implicit Euler errors, though this does not hold for the error in u.

4.2. **Problem 2.** In the second experiment, we investigated the influence of the viscosity on the errors. To this end, we employ the same data as presented in Section 4.1 except for the viscosity operator which we set to

$$\mathbf{A} = \gamma \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

(in Voigt notation). In this case, we used $N_x \in \{4, 8, 16, 32\}$ with $N_t = N_x^2/4$ and took $N_x = 64$, $N_t = 1024$ for the reference approximation. We only used the semi-implicit scheme here. The first observation is that varying γ has essentially no effect on the errors in θ and ϕ. This is to be expected, as the influence of u on θ is not so large. We therefore omit the plots of these errors, and instead present the error in u for different values of γ in Figure 2.

We observe that the error clearly increases as γ is decreased, which is to be expected. Indeed, an inspection of the convergence proof indicates that the L^2-error should be inversely proportional to the coercivity constant of $\mathbf{A}$, and thus also of γ. This is, however, in the worst case. In the current situation, Figure 2

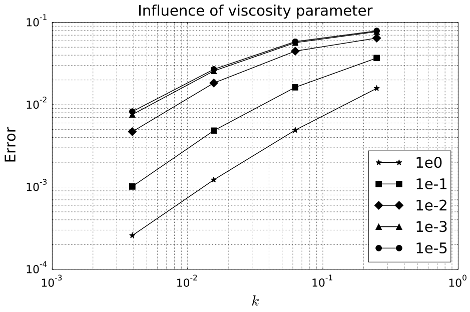

Figure 2. The errors $\max_{1\leq n\leq N_t} \|U_h^n - U_{\text{ref}}^n\|_{L^2}$ for the problem defined in Section 4.2, computed by the semi-implicit method. The different curves correspond to the different values of $\gamma \in \{10^0, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-5}\}$.

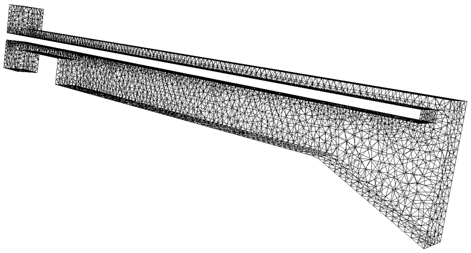

Figure 3. A mesh for the problem described in Section 4.3. The outer dimensions are 192 × 27 × 9.

indicates that even $\gamma = 0$ would be perfectly feasible, though smaller step sizes might be necessary to enter the asymptotic regime.

4.3. **Problem 3.** For our last numerical experiment, we consider a 3D problem arising from an engineering application, inspired by [15] and [16]. We let Ω be as in Figure 3, which also shows a typical spatial tetrahedral discretization. This represents a micro-electro-mechanical system (MEMS) used for precise positioning on small scales. When an electric current is passed through the device from the upper-left connector to the lower-left connector, it heats up. This causes a deformation, which due to the asymmetrical design of the component makes the tip move downwards.

We employ homogeneous Neumann boundary conditions everywhere except for at the left-most edge of the two connectors. These correspond to the component being insulated and stress-free. On the left-most edge we choose the Dirichlet boundary conditions

$$\theta = 0, \quad \phi = \begin{cases} 1, & z > 0 \\ -1, & z < 0 \end{cases}, \quad \text{and} \quad u = v = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

corresponding to the component being clamped and having a potential difference applied between the two connectors. Further, we take $\mathbf{M} = 10^{-1} I$, $f = [0, 0, 0]^T$

h	k	Vertices	Error in θ	Error in ϕ	Error in u
4.84	6.67×10^1	5319	6.18×10^{-2}	4.33×10^{-1}	2.42×10^2
3.73	5.00×10^1	7554	5.40×10^{-2}	2.99×10^{-1}	2.23×10^2
2.80	2.86×10^1	11 888	3.96×10^{-2}	1.76×10^{-1}	1.71×10^2
2.48	2.22×10^1	18 799	3.22×10^{-2}	1.24×10^{-1}	1.42×10^2
2.01	1.54×10^1	28 535	2.12×10^{-2}	8.91×10^{-2}	9.54×10^1
1.33	6.90	85 260	-	-	-

Table 1. Spatial and temporal discretizations parameters as well as maximal errors for the MEMS problem (Section 4.3) at the time points $t_j = 2j \cdot 10^2$ for $j = 1, \ldots, 10$. The last line corresponds to the reference approximation.

and (for simplicity) the viscosity and elasticity operators to be

$$\mathbf{A} = \mathbf{B} \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

in Voigt notation. The electrical conductivity is chosen as in [16],

$$\sigma(\theta) = 0.5 + \frac{\pi}{1.5}\left(\frac{\pi}{2} + \arctan\Big(200(\theta - 0.25)\Big)\right).$$

We solve the problem until the time $T = 2 \cdot 10^3$ using the semi-implicit method for different spatial and temporal discretizations. The maximum sizes h of the tetrahedrons that were used and the corresponding number of vertices are listed in Table 1. The time steps were again taken proportional to h^2, in this case roughly $4h^2$ but modified slightly to yield an integer number of steps. Since the temporal grids thus generated are not refinements of each other, we measured the error as the sum of the errors at only the points $t_j = 2j \cdot 10^2$ for $j = 1, \ldots, 10$. These errors are also listed in Table 1, and plotted in Figure 4. While we cannot apply Theorem 3.3 directly due to the mixed boundary conditions and the non-convexity of the domain, we observe that we still acquire almost $\mathrm{O}(h^2 + k)$ convergence. The different magnitudes of the errors reflect the relative sizes of the solution components.

Finally, Figure 5 shows the approximations Θ_h^N, Φ_h^N and U_h^N at T, viewed from the side. We note that the body deforms in the expected fashion. The chosen T is still in the transient phase before the temperature and deformation have stabilized, and careful inspection shows how the body flexes also in undesirable directions before reaching a steady state. In a real electrical component such deformations might result in unforeseen amounts of material fatigue. This observation therefore provides additional motivation for studying the fully dynamical rather than quasi-static or static process.

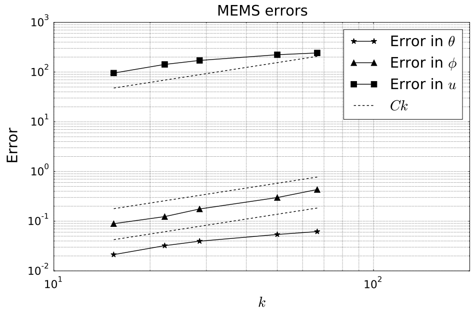

FIGURE 4. Maximal errors at the time points $t_j = 2j \cdot 10^2$ for $j = 1, \dots, 10$ for the MEMS problem defined in Section 4.3.

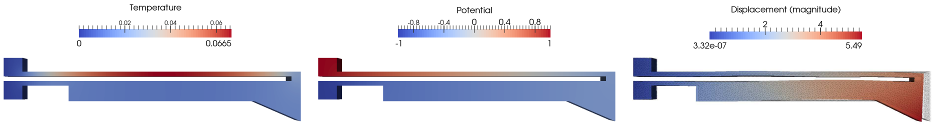

FIGURE 5. The approximation to the solution of the problem defined in Section 4.3 at $t = T$ and with the finest spatial and temporal discretization. In the right-most plot, the grid has been deformed according to the computed displacement and then superimposed over the original mesh to illustrate the deformation. We note that the grid is never deformed in the actual computations. (This figure is in color in the electronic version of the article.)

5. Conclusions and outlook

We have presented a fully discrete numerical method for the fully coupled thermoviscoelastic thermistor problem (1.1)–(1.3) and proved optimal convergence orders in both space and time. These theoretical results are validated by experimental results.

We reiterate that mixed boundary conditions and re-entrant corners might lead to order reductions. In that case an adaptive mesh refinement strategy may be used, which requires a good a posteriori error estimate. It is possible that the ideas in [3] regarding this can be extended to the present, deformable case.

As illustrated by Section 4.3, a typical thermistor is not convex, so a further item that could be improved in the analysis is therefore the shape of the computational domain itself. In this direction we note that the stationary version of the non-deformable problem has been studied in [16, 18] for very general domains. It is our ambition to extend these ideas to the time-dependent deformable case in the future.

References

[1] Akrivis, G., Larsson, S.: Linearly implicit finite element methods for the time-dependent Joule heating problem. BIT **45**(3), 429–442 (2005). DOI 10.1007/s10543-005-0008-1

[2] Allegretto, W., Xie, H.: Existence of solutions for the time-dependent thermistor equations. IMA J. Appl. Math. **48**(3), 271–281 (1992). DOI 10.1093/imamat/48.3.271

[3] Allegretto, W., Yan, N.: A posteriori error analysis for FEM of thermistor problems. Int. J. Numer. Anal. Model. **3**(4), 413–436 (2006)

[4] Alnæs, M.S., Blechta, J., Hake, J., Johansson, A., Kehlet, B., Logg, A., Richardson, C., Ring, J., Rognes, M.E., Wells, G.N.: The FEniCS project version 1.5. Archive of Numerical Software **3**(100) (2015). DOI 10.11588/ans.2015.100.20553

[5] Antontsev, S.N., Chipot, M.: The thermistor problem: existence, smoothness uniqueness, blowup. SIAM J. Math. Anal. **25**(4), 1128–1156 (1994). DOI 10.1137/S0036141092233482

[6] Chen, X.: Existence and regularity of solutions of a nonlinear nonuniformly elliptic system arising from a thermistor problem. J. Partial Differential Equations **7**(1), 19–34 (1994)

[7] Cimatti, G.: Remark on existence and uniqueness for the thermistor problem under mixed boundary conditions. Quart. Appl. Math. **47**(1), 117–121 (1989)

[8] Cimatti, G.: Existence of weak solutions for the nonstationary problem of the Joule heating of a conductor. Ann. Mat. Pura Appl. (4) **162**, 33–42 (1992). DOI 10.1007/BF01759998

[9] Duvaut, G., Lions, J.L.: Inequalities in mechanics and physics. Springer, Berlin (1976)

[10] Elliott, C.M., Larsson, S.: A finite element model for the time-dependent Joule heating problem. Math. Comp. **64**(212), 1433–1453 (1995). DOI 10.2307/2153363

[11] Fernández, J.R.: Numerical analysis of the quasistatic thermoviscoelastic thermistor problem. M2AN Math. Model. Numer. Anal. **40**(2), 353–366 (2006). DOI 10.1051/m2an:2006016

[12] Fernández, J.R., Kuttler, K.L.: A dynamic thermoviscoelastic problem: an existence and uniqueness result. Nonlinear Anal. **72**(11), 4124–4135 (2010). DOI 10.1016/j.na.2010.01.044

[13] Fernández, J.R., Kuttler, K.L.: A dynamic thermoviscoelastic problem: numerical analysis and computational experiments. Quart. J. Mech. Appl. Math. **63**(3), 295–314 (2010). DOI 10.1093/qjmam/hbq012

[14] Grisvard, P.: Elliptic problems in nonsmooth domains, *Monographs and Studies in Mathematics*, vol. 24. Pitman (Advanced Publishing Program), Boston, MA (1985)

[15] Henneken, V.A., Tichem, M., Sarro, P.M.: In-package MEMS-based thermal actuators for micro-assembly. J. Micromech. Microeng. **16**, 107–115 (2006). DOI 10.1088/0960-1317/16/6/S17

[16] Holst, M.J., Larson, M.G., Målqvist, A., Söderlund, R.: Convergence analysis of finite element approximations of the Joule heating problem in three spatial dimensions. BIT **50**(4), 781–795 (2010). DOI 10.1007/s10543-010-0287-z

[17] Howison, S.D., Rodrigues, J.F., Shillor, M.: Stationary solutions to the thermistor problem. J. Math. Anal. Appl. **174**(2), 573–588 (1993). DOI 10.1006/jmaa.1993.1142

[18] Jensen, M., Målqvist, A.: Finite element convergence for the Joule heating problem with mixed boundary conditions. BIT **53**(2), 475–496 (2013)

[19] Kuttler, K.L., Shillor, M., Fernández, J.R.: Existence for the thermoviscoelastic thermistor problem. Differ. Equ. Dyn. Syst. **16**(4), 309–332 (2008). DOI 10.1007/s12591-008-0017-z

[20] Li, B., Gao, H., Sun, W.: Unconditionally optimal error estimates of a Crank-Nicolson Galerkin method for the nonlinear thermistor equations. SIAM J. Numer. Anal. **52**(2), 933–954 (2014). DOI 10.1137/120892465

[21] Li, B., Sun, W.: Error analysis of linearized semi-implicit Galerkin finite element methods for nonlinear parabolic equations. Int. J. Numer. Anal. Model. **10**(3), 622–633 (2013)

[22] Lin, Y.P., Thomée, V., Wahlbin, L.B.: Ritz-Volterra projections to finite-element spaces and applications to integrodifferential and related equations. SIAM J. Numer. Anal. **28**(4), 1047–1070 (1991). DOI 10.1137/0728056

[23] Logg, A., Mardal, K.A., Wells, G.N., et al.: Automated Solution of Differential Equations by the Finite Element Method. Springer, Berlin (2012). DOI 10.1007/978-3-642-23099-8

[24] Nitsche, J.A.: On Korn's second inequality. RAIRO Anal. Numér. **15**(3), 237–248 (1981)

[25] Thomée, V., Zhang, N.Y.: Error estimates for semidiscrete finite element methods for parabolic integro-differential equations. Math. Comp. **53**(187), 121–139 (1989). DOI 10.2307/2008352

[26] Wu, X., Xu, X.: Existence for the thermoelastic thermistor problem. J. Math. Anal. Appl. **319**(1), 124–138 (2006). DOI 10.1016/j.jmaa.2006.01.076

[27] Yuan, G.W., Liu, Z.H.: Existence and uniqueness of the C^α solution for the thermistor problem with mixed boundary value. SIAM J. Math. Anal. **25**(4), 1157–1166 (1994). DOI 10.1137/S0036141092237893

Mathematical Sciences, Chalmers University of Technology and the University of Gothenburg, , SE-412 96 Göteborg, Sweden.
E-mail address, A. Målqvist: axel@chalmers.se

E-mail address, T. Stillfjord: tony.stillfjord@gu.se