Thesis for the Degree of Licentiate of Philosophy

A generalized finite element method for linear thermoelasticity

Anna Persson

Division of Mathematics
Department of Mathematical Sciences
Chalmers University of Technology
and University of Gothenburg
Gothenburg, Sweden, 2016

A generalized finite element method for linear thermoelasticity
Anna Persson

Department of Mathematical Sciences
Chalmers University of Technology
and University of Gothenburg
SE–412 96 Gothenburg
Sweden
Phone: +46 (0)31-772 10 00

Printed in Gothenburg, Sweden, 2016.

Abstract

In this thesis we develop a generalized finite element method for linear thermoelasticity problems, modeling displacement and temperature in an elastic body. We focus on strongly heterogeneous materials, like composites. For classical finite element methods such problems are known to be numerically challenging due to the rapid variations in the data.

The method we propose is based on the local orthogonal decomposition technique introduced in [12]. In short, the idea is to enrich the classical finite element nodal basis function using information from the diffusion coefficient. Locally, these basis functions have better approximation properties than the nodal basis functions.

The papers included in this thesis first extends the local orthogonal decomposition framework to parabolic problems (Paper I) and to linear elasticity equations (Paper II). Finally, using the theory developed in these papers, we address the linear thermoelastic system (Paper III).

Keywords: Thermoelasticity, parabolic equations, linear elasticity, multiscale, composites, generalized finite element, local orthogonal decomposition, a priori analysis.

List of included papers

The following papers are included in this thesis:

- **Paper I.** Axel Målqvist and Anna Persson: *Multiscale techniques for parabolic equations*, Submitted.

- **Paper II.** Patrick Henning and Anna Persson: *A multiscale method for linear elasticity reducing Poisson locking*, Submitted.

- **Paper III.** Axel Målqvist and Anna Persson: *A generalized finite element method for linear thermoelasticity*, Submitted.

Acknowledgments

First and foremost, I would like to thank my supervisor Axel Målqvist for his guidance throughout my work on this thesis. Thank you for introducing me to the subject, for sharing your expertise and knowledge, and for teaching me how the academic world works. I would also like to thank my co-advisors Patrick Henning and Stig Larsson for many fruitful discussions and your helpful advice. Especially, thank you Patrick for the last months of collaboration and for always being so positive.

Furthermore, I am grateful to all my colleagues at the mathematics department for creating such a friendly working environment. Many thanks also goes to my family for always believing in me and giving me encouragement when I need it.

Last, but not least, thank you Joakim for your love and immense support.

Contents

Part 1

Introduction

Introduction

1. Background

In many applications the expansion and contraction of a material exposed to external forces and temperature changes are of great importance. For instance, it may be crucial when designing parts for aircrafts or when constructing a bridge.

In this thesis we study numerical solutions to linear thermoelastic systems, which consist of partial differential equations (PDEs) simulating displacement and temperature changes in materials over time. In particular, we are interested in applications where the material under consideration is strongly heterogeneous, e.g. composites. Composite materials are constructed using two or more different constituents. Typically, the material properties in composites vary on a very fine scale, as in, for instance, fiber reinforced materials. Modeling physical behavior in these materials results in equations with highly varying and oscillating coefficients. Such problems, that exhibit a lot of variations in the data, often on multiple scales, are commonly referred to as *multiscale* problems.

Classically, numerical solutions to thermoelasticity equations have been obtained using finite element methods (FEMs) based on continuous piecewise polynomials. These methods work well for homogeneous materials, or materials that are not varying too much in space. However, for strongly heterogeneous materials the classical FEMs struggle to approximate the solution accurately unless the mesh width is sufficiently small. Indeed, the mesh width must be small enough to resolve all the fine variations in the data. In practice, this leads to issues with computational cost and available memory.

Today's increasing interest in and usage of composite materials thus pose a demand for other types of numerical methods. Several such methods have been proposed over the last two decades, see, for instance, [9, 5, 1, 10]. However, the analysis of many of these methods require restrictive assumptions on the material, such as periodicity or separation of scales.

In [12] a generalized finite element method (GFEM), cf. [2], is proposed and rigorous analysis is provided. Convergence of the method is proven for an arbitrary positive and bounded coefficient, that is, no assumptions on periodicity or separation of scales are needed.

The purpose of this thesis is to generalize the method proposed in [12] to solve linear thermoelasticity equations with highly varying and oscillating coefficients. This is done in three steps. In Paper I we extend the method to linear parabolic problems, in Paper II we consider (stationary) linear elasticity equations and in Paper III we finally address the thermoelastic system. In all three papers we prove convergence of optimal order for highly varying coefficients and provide several numerical examples that confirm the analysis.

In the upcoming section we describe the system of equations used to model the displacement and temperature of an elastic material. In Section 2 the issue with applying the classical FEM to multiscale problems is described in more detail. In Section 3 we introduce the GFEM proposed in [12] for elliptic equations and discuss the main idea behind the extension to linear thermoelasticity. Finally, in Section 4 we summarize the appended papers and highlight the main results.

1.1. Linear thermoelasticity. Linear thermoelasticity refers to a coupled system of PDEs describing the displacement and temperature of an elastic body, see [3, 4]. To introduce the mathematical formulation of this system we let $\Omega \subseteq \mathbb{R}^d$, $d = 1, 2, 3$, be a domain describing the initial configuration of an elastic medium. For a given simulation time $T > 0$, we let the vector valued function $u : [0, T] \times \Omega \to \mathbb{R}^d$ denote the displacement field and $\theta : [0, T] \times \Omega \to \mathbb{R}$ denote the temperature. To define boundary conditions for u we let Γ^u_D and Γ^u_N be two disjoint parts of the boundary such that $\Gamma^u_D \cup \Gamma^u_N = \partial\Omega$. On the part denoted Γ^u_D we impose Dirichlet boundary conditions corresponding to a clamped part of the material. On Γ^u_N, corresponding to the traction boundary, we impose Neumann boundary conditions. Similarly, we define Γ^θ_D and Γ^θ_N to be the drained and flux part of the boundary for the temperature θ.

Under the assumption that the displacement gradients are small, the strain tensor is given by the following linear relation

$$\varepsilon(u) = \frac{1}{2}(\nabla u + \nabla u^{\intercal}).$$

For isotropic materials the total stress tensor is given by

$$\bar{\sigma} = 2\mu\varepsilon(u) + \lambda(\nabla \cdot u)I - \alpha\theta I,$$

where I is the d-dimensional identity matrix and α is the thermal expansion coefficient. Furthermore, μ and λ denotes the Lamé coefficients satisfying

$$\mu = \frac{E}{2(1+\nu)}, \quad \lambda = \frac{E\nu}{(1+\nu)(1-2\nu)},$$

where ν denotes Poisson's ratio and E denotes Young's elastic modulus. Poisson's ratio is a measure on the materials tendency to shrink (expand) when stretched (compressed) and Young's modulus describes the stiffness of the material. The coefficients α, λ, and μ are all material dependent and thus rapidly varying in space for strongly heterogeneous (multiscale) materials.

Now, Cauchy's equilibrium equations states that

$$-\nabla \cdot \bar{\sigma} = f,$$

where $f : \Omega \to \mathbb{R}^d$ denotes the external body forces. Furthermore, the temperature in the material can be described by the parabolic equation

$$\dot{\theta} - \nabla \cdot \kappa \nabla \theta + \alpha \nabla \cdot \dot{u} = g,$$

where $\kappa : \Omega \to \mathbb{R}^{d \times d}$ is the heat conductivity parameter and g denotes internal heat sources. Note that κ is material dependent and thus rapidly varying. To summarize, the linear thermoelastic system is given by the following system of equations

$$\begin{aligned}
&(1.1) & -\nabla \cdot (2\mu\varepsilon(u) + \lambda \nabla \cdot uI - \alpha\theta I) &= f, && \text{in } (0,T] \times \Omega, \\
&(1.2) & \dot{\theta} - \nabla \cdot \kappa \nabla \theta + \alpha \nabla \cdot \dot{u} &= g, && \text{in } (0,T] \times \Omega, \\
&(1.3) & u &= 0, && \text{in } (0,T] \times \Gamma_D^u, \\
&(1.4) & \bar{\sigma} \cdot n &= 0, && \text{in } (0,T] \times \Gamma_N^u. \\
&(1.5) & \theta &= 0, && \text{on } (0,T] \times \Gamma_D^\theta, \\
&(1.6) & \kappa \nabla \theta \cdot n &= 0, && \text{on } (0,T] \times \Gamma_N^\theta. \\
&(1.7) & \theta(0) &= \theta_0, && \text{in } \Omega,
\end{aligned}$$

where we for simplicity assume homogeneous boundary conditions. Note that the equations (1.1)-(1.2) are coupled.

Remark 1.1. The system (1.1)-(1.7) is formally equivalent to a linear model for poroelasticity. In this case θ denotes the fluid pressure, κ the hydraulic conductivity, and α the Biot-Willis coupling-deformation coefficient. Hence, the results in this thesis also apply to the linear poroelastic system.

To define a FEM (and a GFEM) for (1.1)-(1.7) we define the corresponding variational (or weak) formulation. For this purpose we first need to introduce some notation and spaces. We use $(\cdot,\cdot)$ to denote the inner product in $L_2(\Omega)$ and $\|\cdot\|$ the corresponding norm. Let $H^1(\Omega) := W_2^1(\Omega)$ denote the classical Sobolev space with norm $\|v\|_{H^1(\Omega)}^2 = \|v\|^2 + \|\nabla v\|^2$ and let $H^{-1}(\Omega)$ denote the dual space to H^1. Furthermore, let $L_p([0,T];X)$ denote the Bochner space with norm

$$\begin{aligned}
\|v\|_{L_p([0,T];X)} &= \Big(\int_0^T \|v\|_X^p \,\mathrm{dt} \Big)^{1/p}, \quad 1 \le p < \infty, \\
\|v\|_{L_\infty([0,T];X)} &= \operatorname*{ess\,sup}_{0 \le t \le T} \|v\|_X,
\end{aligned}$$

where X is a Banach space equipped with the norm $\|\cdot\|_X$. The dependence on the interval $[0,T]$ and the domain Ω is frequently suppressed and we write, for instance, $L_2(L_2)$ for $L_2([0,T];L_2(\Omega))$. We also use the double-dot product notation to denote the Frobenius inner product of two matrices A and B

$$A : B = \sum_{i,j=1}^{d} A_{ij} B_{ij}, \quad A, B \in \mathbb{R}^{d\times d}.$$

Now, define the following spaces

$$V^1 := \{v \in (H^1(\Omega))^d : v = 0 \text{ on } \Gamma_D^u\}, \quad V^2 := \{v \in H^1(\Omega) : v = 0 \text{ on } \Gamma_D^\theta\}.$$

Multiplying (1.1) with $v_1 \in V^1$ and (1.2) with $v_2 \in V^2$ and using Green's formula together with the boundary conditions (1.3)-(1.6) we arrive at the following variational formulation; find $u(t,\cdot) \in V^1$ and $\theta(t,\cdot) \in V^2$ such that, for a. e. $t > 0$,

$$(\sigma(u) : \varepsilon(v_1)) - (\alpha\theta, \nabla \cdot v_1) = (f, v_1), \quad \forall v_1 \in V^1, \tag{1.8}$$

$$(\dot{\theta}, v_2) + (\kappa\nabla\theta, \nabla v_2) + (\alpha\nabla \cdot \dot{u}, v_2) = (g, v_2), \quad \forall v_2 \in V^2, \tag{1.9}$$

and the initial value $\theta(0,\cdot) = \theta_0$ is satisfied. Here $\sigma(u) := 2\mu\varepsilon(u) + \lambda\nabla \cdot uI$ is the first part of $\bar{\sigma}$ involving only the displacement u, commonly referred to as the effective stress tensor.

Two functions u and θ are weak solutions if (1.8)-(1.9) are satisfied and $u \in L_2(V^1)$, $\nabla \cdot \dot{u} \in L_2(H^{-1})$, $\theta \in L_2(V^2)$, and $\dot{\theta} \in L_2(H^{-1})$. Existence and uniqueness of such weak solutions are proved in, e.g., [17, 16], and in [14] within the framework of linear degenerate evolution equations in Hilbert spaces. In [14] it is also proved that the system is of parabolic type, meaning that it is well posed for nonsmooth initial data with regularity estimates depending on negative powers of t.

2. Classical finite element

In this section we explain more carefully why the classical FEM fails to approximate the solution to problems with rapidly varying data. To simplify the discussion we start by considering elliptic equations.

2.1. Elliptic equations. Consider the elliptic equation

$$\begin{aligned} -\nabla \cdot A\nabla u &= f, \quad \text{in } \Omega, \\ u &= 0, \quad \text{on } \partial\Omega, \end{aligned}$$

with the variational formulation; find $u \in V$, such that

$$a(u, v) = (f, v), \quad \forall v \in V, \tag{2.1}$$

where $V = H_0^1(\Omega)$ and $a(u, v) := (A\nabla u, \nabla v)$. Here $A : \Omega \to \mathbb{R}^{d\times d}$ the diffusion coefficient is assumed to be rapidly oscillating.

To define a FEM we need a triangulation of the domain. Let $\{\mathcal{T}_h\}_{h>0}$ be a family triangulations of Ω with the mesh size $h_K := \mathrm{diam}(K)$, for $K \in \mathcal{T}_h$ and denote the largest diameter in the triangulation by $h := \max_{K\in\mathcal{T}_h} h_K$. Now let $V_h \subseteq V$ denote the space of continuous piecewise affine functions on the triangulation $\mathcal{T}_h$. The finite element formulation then reads; find $u_h \in V_h$, such that,

$$a(u_h, v) = (f, v), \quad \forall v \in V_h. \tag{2.2}$$

Classical a priori error analysis gives the bound

$$\|u_h - u\|_{H^1} \leq Ch\|D^2 u\|, \tag{2.3}$$

where $D^2 u$ denotes the second order (weak) derivatives of u. Not only does this bound require additional regularity of the solution, the norm $\|D^2 u\|$ may also be very large if A is rapidly oscillating. Indeed, if A varies with frequency ϵ^{-1} for some $\epsilon > 0$, then $\|\nabla A\|_{L_\infty} = O(\epsilon^{-1})$. Estimating $\|D^2 u\|$ with the problem data gives

$$\begin{aligned} \|D^2 u\| \leq C\|\Delta u\| &\leq C\|A\nabla \cdot (\nabla u)\| = C\|\nabla \cdot (A\nabla u) - \nabla A \nabla u\| \\ &\leq C\|\nabla \cdot (A\nabla u)\| + \|\nabla A\|_{L_\infty}\|\nabla u\| \leq C(1 + \epsilon^{-1})\|f\|, \end{aligned}$$

where we used elliptic regularity in the first inequality and the bound $\|u\|_{H^1} \leq C\|f\|$, derived from (2.1), in the last inequality. Furthermore, we can derive the bound $\|u_h\|_{H^1} \leq C\|f\|$ from (2.2), which gives the following upper bound of the error; $\|u_h - u\|_{H^1} \leq C\|f\|$. Hence, the error bound (2.3) takes the form

$$\|u_h - u\|_{H^1} \leq C \min\left\{h + \frac{h}{\epsilon}, 1\right\} \|f\|,$$

and convergence does not take place unless $h < \epsilon$. If ϵ is small, the condition $h < \epsilon$, can be devastating considering computational cost and available memory.

2.2. Linear thermoelasticity. As in the previous section we define a family of triangulations $\{\mathcal{T}_h\}_{h>0}$ and we let $V_h^1 \subseteq V^1$ and $V_h^2 \subseteq V^2$ denote finite element spaces consisting of continuous piecewise linear functions on this triangulation. Furthermore, we let $0 = t_0 < t_1 < ... < t_N = T$ be a uniform discretization of the time interval such that $t_j - t_{j-1} = \tau > 0$ for $j = 1, ..., N$. The classical FEM with a backward (implicit) Euler discretization in time for (1.8)-(1.9) reads; for $n \in \{1, ..., N\}$ find $u_h^n \in V_h^1$ and $\theta_h^n \in V_h^2$, such that

$$(\sigma(u_h^n) : \varepsilon(v_1)) - (\alpha\theta_h^n, \nabla \cdot v_1) = (f^n, v_1), \quad \forall v_1 \in V_h^1, \tag{2.4}$$

$$(\bar{\partial}_t \theta_h^n, v_2) + (\kappa \nabla \theta_h^n, \nabla v_2) + (\alpha \nabla \cdot \bar{\partial}_t u_h^n, v_2) = (g^n, v_2), \quad \forall v_2 \in V_h^2, \tag{2.5}$$

where $\bar{\partial}_t \theta_h^n := (\theta_h^n - \theta_h^{n-1})/\tau$ and similarly for $\bar{\partial}_t u_h^n$. Here $u_h^0 = u_{h,0}$ and $\theta_h^0 = \theta_{h,0}$, where $u_{h,0} \in V_h^1$ and $\theta_{h,0} \in V_h^2$ denote suitable initial conditions. The right hand sides are evaluated at time t_n, that is, $f^n := f(t_n)$ and $g^n := g(t_n)$.

A priori analysis for the system (2.4)-(2.5) can be found in [7]. It follows that the error is bounded by

$$\|u_h^n - u^n\|_{H^1} + \|\theta_h^n - \theta^n\| + \left(\sum_{j=1}^n \tau\|\theta_h^j - \theta^j\|_{H^1}^2\right)^{1/2} \leq C_{\epsilon^{-1}} h + C\tau,$$

where the constant $C_{\epsilon^{-1}}$ depends on both $\|u(t_n)\|_{H^2}$ and $\|\theta(t_n)\|_{H^2}$. By arguments similar to the ones used for the elliptic equation in Section 2.1, we get that

$\|u(t_n)\|_{H^2} = O(\epsilon^{-1})$ and $\|\theta(t_n)\|_{H^2} = O(\epsilon^{-1})$, if the material has variations on a scale of size ϵ.

3. A generalized finite element method

In [12] a GFEM, often referred to as *local orthogonal decomposition*, is proposed and analyzed for elliptic equations of the form (2.1). In Section 3.1 below we describe this method and the main ideas used in the analysis. Finally, in Section 3.4 we describe how this method can be generalized to define a GFEM for linear thermoelasticity, which is the main objective of this thesis.

3.1. Elliptic equations.

The method proposed in [12] builds on the ideas from the variational multiscale method [10, 11], where the solution space is decomposed to into a coarse and a fine part. The nodal basis functions in the coarse space is then modified by adding a *correction* from the fine space.

We begin by assuming that the mesh size h used in the classical FEM in (2.2) is fix and sufficiently small, that is $h < \epsilon$, such that the error (2.3) is small. In this case, the solution u_h and the space V_h are referred to as the reference solution and the reference space, respectively. Now define V_H similarly to V_h but with a larger mesh size $H > h$, such that $V_H \subseteq V_h$. Note that the classical FEM solution u_H in the coarse space V_H is not a good approximation to u. It is, however, cheaper to compute than u_h since $\dim(V_H) < \dim(V_h)$. The aim is now to define a new multiscale space V_{ms} with the same dimension as the coarse space V_H, but with better approximations properties.

To define such a space, we need an interpolation operator $I_H : V_h \to V_H$ with the properties $I_H \circ I_H = I_H$ and for $K \in \mathcal{T}_H$

$$H_K^{-1}\|v - I_H v\|_{L_2(K)} + \|\nabla I_H v\|_{L_2(K)} \leq C_I \|\nabla v\|_{L_2(\omega_K)}, \quad v \in V_h, \tag{3.1}$$

where $\omega_K := \cup\{\hat{K} \in \mathcal{T}_H : \hat{K} \cap K \neq \emptyset\}$. For a quasi-uniform mesh, the bounds in (3.1) can be summed to achieve a global bound

$$H^{-1}\|v - I_H v\| + \|\nabla I_H v\| \leq C\|\nabla v\|, \tag{3.2}$$

There are many interpolations operators that satisfy these conditions, for instance, the global L_2-projection. In Paper II and Paper III we use an interpolation of the form $I_H = E_H \circ \Pi_H$, where Π_H is the L_2-projection onto $P_1(\mathcal{T}_H)$, the space of functions that are affine on each triangle $K \in \mathcal{T}_H$ and $E_H : P_1(\mathcal{T}_H) \to V_H$ is an averaging operator. We refer to [13, 6] for further details and possible choices of I_H.

Now let V_{f} denote the kernel to the operator I_H

$$V_{\mathrm{f}} := \ker I_H = \{v \in V_h : I_H v = 0\}.$$

The space V_h can be decomposed as $V_h = V_H \oplus V_{\mathrm{f}}$, meaning that $v_h \in V_h$ can be decomposed into

$$v_h = v_H + v_{\mathrm{f}}, \quad v_H \in V_H, \; v_{\mathrm{f}} \in V_{\mathrm{f}}. \tag{3.3}$$

The kernel V_{f} is a fine scale (detail) space in the sense that it captures all features that are not captured by the coarse space V_H. Let $R_{\mathrm{f}} : V_h \to V_{\mathrm{f}}$ denote the Ritz projection onto V_{f}, that is,

$$a(R_{\mathrm{f}}v, w) = a(v, w), \quad \forall w \in V_{\mathrm{f}},\ v \in V_h. \tag{3.4}$$

Because of the decomposition (3.3) we have the identity

$$v_h - R_{\mathrm{f}}v_h = v_H + v_{\mathrm{f}} + R_{\mathrm{f}}(v_H + v_{\mathrm{f}}) = v_H - R_{\mathrm{f}}v_H,$$

since $v_{\mathrm{f}} \in V_{\mathrm{f}}$. Using this we can define the multiscale space V_{ms}

$$V_{\mathrm{ms}} := V_h - R_{\mathrm{f}}V_h = V_H - R_{\mathrm{f}}V_H. \tag{3.5}$$

Note that V_{ms} is the orthogonal complement to V_{f} with respect to the inner product $a(\cdot,\cdot)$ and must have the same dimension as V_H. Indeed, with $\mathcal{N}$ denoting the inner nodes in $\mathcal{T}_H$ and λ_z the basis function at node z, a basis for V_{ms} is given by

$$\{z \in \mathcal{N} : \lambda_z - R_{\mathrm{f}}\lambda_z\}.$$

Hence, that basis functions are the classical nodal basis functions modified by corrections $R_{\mathrm{f}}\lambda_z$ computed in the fine scale space.

Replacing V_h with V_{ms} in (2.2) we can now define the GFEM; find $u_{\mathrm{ms}} \in V_{\mathrm{ms}}$, such that,

$$a(u_{\mathrm{ms}}, v) = (f, v), \quad \forall v \in V_{\mathrm{ms}}. \tag{3.6}$$

The following theorem gives an a priori bound for the GFEM and can be found in [12]. We include the proof here since it is short and highlights the main ideas used in the analysis.

Theorem 3.1. *Let u_h be the solution to* (2.2) *and u_{ms} the solution to* (3.6). *Then*

$$\|u_{\mathrm{ms}} - u_h\|_{H^1} \leq CH\|f\|,$$

where C does not depend on the derivatives of A.

Proof. Define $e := u_{\mathrm{ms}} - u_h$ and note that $e \in V_{\mathrm{f}}$. Hence, $I_He = 0$. Furthermore we have due to Galerkin orthogonality $a(e, v_{\mathrm{ms}}) = 0$ for $v_{\mathrm{ms}} \in V_{\mathrm{ms}}$. Using this together with the interpolation bound (3.2) we have

$$a(e,e) = -a(e,u_h) = -(f,e) \leq \|f\|\|e\| = \|f\|\|e - I_He\| \leq CH\|f\|\|\nabla e\|,$$

and the bound follows by using equivalence of the energy norm induced by $a(\cdot,\cdot)$ and the H^1-norm. □

From Theorem 3.1 we have that the solution given by the GFEM converges to u_h, with optimal order, independently of the derivatives (variations) of A. We emphasize that the total error is bounded by

$$\|u_{\mathrm{ms}} - u\|_{H^1} \leq \|u_{\mathrm{ms}} - u_h\|_{H^1} + \|u_h - u\|_{H^1},$$

where the error in the second term is due to the classical FEM and assumed to be of reasonable size, since h is assumed to be sufficiently small.

Although the a priori analysis seems promising, the GFEM as suggested above suffers from some drawbacks. The problem of finding the corrections $R_{\mathrm{f}}\lambda_z$, which are needed to construct the basis, are posed in the entire fine scale space V_{f} which has the same dimension as V_h. Furthermore, the corrections generally have global support and therefore destroys the sparsity of the resulting discrete system. Both issues are resolved by performing a localization of the corrections. The localization is motivated by the observation that the correction $R_{\mathrm{f}}\lambda_z$ decay exponentially away from node z.

3.2. Localization. In [12] it is proved that the corrections decay exponentially and a localization procedure is proposed. However, in [8] a different localization technique is proposed which allows for smaller patches to be used. We describe the procedure in [8] here, which is also the procedure that is used in the appended papers.

We define patches of size k in the following way; for $K \in \mathcal{T}_H$

$$\begin{aligned}
&\omega_0(K) := \text{int } K,\\
&\omega_k(K) := \text{int } \big(\cup \{\hat{K} \in \mathcal{T}_H : \hat{K} \cap \overline{\omega_{k-1}(K)} \neq \emptyset\}\big), \quad k = 1, 2, ...,
\end{aligned}$$

and let $V_{\mathrm{f}}(\omega_k(K)) := \{v \in V_{\mathrm{f}} : v(z) = 0 \text{ on } \overline{\Omega} \setminus \omega_k(K)\}$ be the restriction of V_{f} to the patch $\omega_k(K)$.

We proceed by noting that R_{f} in (3.4) can be written as the sum

$$R_{\mathrm{f}} = \sum_{K \in \mathcal{T}_H} R_{\mathrm{f}}^K,$$

where $R_{\mathrm{f}}^K : V_h \to V_{\mathrm{f}}$ and fulfills

$$a(R_{\mathrm{f}}^K v, w) = a(v, w)_K, \quad \forall w \in V_{\mathrm{f}}, \quad v \in V_h, \quad K \in \mathcal{T}_H, \tag{3.7}$$

where we define

$$a(v, w)_K := (A\nabla v, \nabla w)_{L_2(K)}, \quad K \in \mathcal{T}_H.$$

The aim is to localize these computations by replacing V_{f} with $V_{\mathrm{f}}(\omega_k(K))$. Define $R_{\mathrm{f},k}^K : V_h \to V_{\mathrm{f}}(\omega_k(K))$ such that

$$a(R_{\mathrm{f},k}^K v, w) = a(v, w)_K, \quad \forall w \in V_{\mathrm{f}}(\omega_k(K)), \quad v \in V_h, \quad K \in \mathcal{T}_H,$$

and set $R_{\mathrm{f},k} := \sum_{K \in \mathcal{T}_H} R_{\mathrm{f},k}^K$. We can now define the localized multiscale space

$$V_{\mathrm{ms},k} = \{v_H - R_{\mathrm{f},k} v_H : v_H \in V_H\}. \tag{3.8}$$

By replacing V_{ms} with $V_{\mathrm{ms},k}$ in (3.6) a localized GFEM can be defined; find $u_{\mathrm{ms},k} \in V_{\mathrm{ms},k}$ such that

$$a(u_{\mathrm{ms},k}, v) = (f, v), \quad \forall v \in V_{\mathrm{ms},k}. \tag{3.9}$$

Since the dimension of $V_{\mathrm{f}}(\omega_k(K))$ can be made significantly smaller than the dimension of V_{f} (depending on k), the problem of finding $R_{\mathrm{f},k}\lambda_z$ is computationally cheaper than finding $R_{\mathrm{f}}\lambda_z$. Moreover, the resulting discrete system is sparse. It should also be noted that the computation of $R_{\mathrm{f},k}\lambda_z$ for all nodes z is suitable for parallelization, since they are independent of each other.

The convergence of the method (3.9) depends on the size of the patches. In [12, 8] the following Theorem is proved.

THEOREM 3.2. *Let u_h be the solution to* (2.2) *and $u_{\mathrm{ms},k}$ the solution to* (3.9)*. Then there exists $\xi \in (0,1)$ such that*

$$\|u_{\mathrm{ms},k} - u_h\|_{H^1} \leq C(H + k^{d/2}\xi^k)\|f\|,$$

where C does not depend on the derivatives of A.

To achieve linear convergence k should be chosen proportional to $\log H^{-1}$, that is, $k = c\log H^{-1}$, for some constant c.

3.3. Parabolic equations. A natural first step in generalizing the GFEM to linear thermoelasticity is to first extend it to a time dependent problem of parabolic type. Recall that the thermoelastic system (1.8)-(1.9) is parabolic [14]. This is the subject of Paper I.

We consider a parabolic problem on the following weak form; find $u(t) \in V$, such that, $u(0) = u_0$ and

$$(\dot{u}, v) + a(u, v) = (f, v), \quad \forall v \in V, \tag{3.10}$$

where $a(u,v) = (A\nabla u, \nabla v)$ as in the elliptic equation (2.1). The diffusion coefficient $A : \Omega \to \mathbb{R}^{d\times d}$ is assumed to not depend on time.

The classical FEM for (3.10) with a backward Euler discretization reads; for $n \in \{1, ..., N\}$ find $u_h^n \in V_h$, such that, $u_h^0 = u_{h,0}$

$$(\bar{\partial}_t u_h^n, v) + a(u_h^n, v) = (f^n, v), \quad \forall v \in V_h, \tag{3.11}$$

with the notation and time discretization as in Section 2.2 and $u_{h,0}$ a suitable approximation of u_0. It is well known, see, e.g., [15], that the following error estimate holds for the parabolic equation

$$\|u_h^n - u(t_n)\|_{H^1} \leq C_{\epsilon^{-1}}h + C\tau,$$

where $C_{\epsilon^{-1}}$ is constant depending on, among other terms, $\|u(t_n)\|_{H^2}$ and is thus of size ϵ^{-1} if A varies on scale of size ϵ. Hence, parabolic problems suffers from the same issues as elliptic problems when using classical finite element.

In the error analysis of the classical FEM, the error is usually split into the two parts

$$u_h^n - u(t_n) = u_h^n - R_h u(t_n) + R_h u(t_n) - u(t_n) =: \theta^n + \rho^n,$$

where $R_h : V \to V_h$ is the Ritz projection given by

$$a(R_h v, w) = a(v, w), \quad \forall w \in V_h, \ v \in V.$$

The error of the Ritz projection is given by the analysis of the elliptic problem

$$\|R_h v - v\|_{H^1} \leq Ch\|D^2 v\|. \tag{3.12}$$

This directly gives the error of ρ^n. Indeed, $\|\rho^n\|_{H^1} \leq Ch\|D^2 u(t_n)\|$, where $\|D^2 u(t_n)\| \leq C_{\epsilon^{-1}}\|\nabla \cdot A\nabla u(t_n)\| = C_{\epsilon^{-1}}\|f^n - \dot{u}(t_n)\|$ and $C_{\epsilon^{-1}}$ depend on the derivatives of A. Furthermore, to bound $\|\theta^n\|_{H^1}$ we put θ^n into (3.11), which gives

$$\begin{aligned}(\bar{\partial}_t\theta^n, v) + a(\theta^n, v) &= -((R_h - I)\bar{\partial}_t u(t_n) + (\bar{\partial}_t u(t_n) - \dot{u}(t_n)), v) \\ &=: -(\bar{\partial}_t\rho^n + \omega, v),\end{aligned}$$

where the error of $\bar{\partial}_t\rho^n$ follows from (3.12) and the error of ω follows from Taylor's formula. In order to bound θ^n in the H^1-norm we can choose $v = \bar{\partial}_t\theta^n$.

Inspired by this we propose the following GFEM for the parabolic problem, where the space V_h in (3.11) is simply replaced by the multiscale space V_{ms} defined in Section 3.1; for $n \in \{1, ..., N\}$ find $u^n_{\text{ms}} \in V_{\text{ms}}$, such that, $u^0_{\text{ms}} = u_{\text{ms},0}$

$$(\bar{\partial}_t u^n_{\text{ms}}, v) + a(u^n_{\text{ms}}, v) = (f^n, v), \quad \forall v \in V_{\text{ms}}, \tag{3.13}$$

with $u_{\text{ms},0}$ a suitable approximation of $u_{h,0}$. Now, because of the choice of the space V_{ms} we can define a Ritz projection $R_{\text{ms}} : V_h \to V_{\text{ms}}$ by

$$a(R_{\text{ms}}v, w) = a(v, w) = (\mathcal{A}_h v, w), \quad \forall v \in V_{\text{ms}},$$

where $\mathcal{A}_h : V_h \to V_h$ is the operator defined by

$$(\mathcal{A}_h v, w) = a(v, w), \quad \forall w \in V_h.$$

The error analysis for the elliptic problem in [12] gives the bound

$$\|R_{\text{ms}}v - v\|_{H^1} \leq CH\|\mathcal{A}_h v\|, \quad \forall v \in V_h, \tag{3.14}$$

where C is independent of the derivatives of A. The assumption that A does not depend on time is crucial here. Otherwise, we would have to define a new space and compute a new set of basis functions at each time step t_n.

As for the elliptic equation we assume that h is sufficiently small to resolve the variations in A. This means that the reference solution u_h given by (3.11) approximates u in (3.10) sufficiently well. In the error analysis we can thus split

$$\|u^n_{\text{ms}} - u(t_n)\|_{H^1} \leq \|u^n_{\text{ms}} - u^n_h\|_{H^1} + \|u^n_h - u(t_n)\|_{H^1},$$

where the second part is bounded by classical FEM error analysis. For the first part we can use a similar analysis, but with the new Ritz projection R_{ms}. We split the error into the parts

$$u^n_{\text{ms}} - u^n_h = u^n_{\text{ms}} - R_{\text{ms}}u^n_h + R_{\text{ms}}u^n_h - u^n_h =: \theta^n_{\text{ms}} + \rho^n_{\text{ms}},$$

where the error of ρ^n_{ms} is given by (3.14) and $\mathcal{A}_h u^n_h = P_h f^n - \bar{\partial}_t u^n_h$ with P_h denoting the L_2-projection onto V_h. For θ^n_{ms} we get by plugging θ^n_{ms} into (3.13)

$$(\bar{\partial}_t\theta^n_{\text{ms}}, v) + a(\theta^n_{\text{ms}}, v) = -(\bar{\partial}_t\rho^n_{\text{ms}}, v), \quad \forall v \in V_{\text{ms}}.$$

Naturally, the error bound in this case depends on the regularity of the (discrete) time derivative of the reference solution. Since the initial data is not in H^2 we expect, for instance, $\|\bar{\partial}_t u^n_h\|$ to depend on negative powers of t_n. This is possible since the backward Euler scheme preserves the smoothing effect of parabolic problems. In Paper I this is thoroughly investigated and error bounds involving negative powers of t_n are derived.

To utilize the localization introduced in Section 3.1 we can replace V_{ms} by $V_{\mathrm{ms},k}$, define a new Ritz projection $R_{\mathrm{ms},k} : V_h \to V_{\mathrm{ms},k}$, and perform similar splits of the error.

3.4. Linear thermoelasticity. In the classical finite element error analysis for linear thermoelasticity, a Ritz projection related to the stationary form of the problem is used to split the error into two terms. This Ritz projection is defined by the following $R_h(v_1, v_2) : V^1 \times V^2 \to V_h^1 \times V_h^2$, such that, $R_h(v_1, v_2) = (R_h^1(v_1, v_2), R_h^2 v_2)$ and for all $(v_1, v_2) \in V^1 \times V^2$,

$$\begin{aligned}(\sigma(v_1 - R_h^1(v_1, v_2)) : \varepsilon(w_1)) - (\alpha(v_2 - R_h^2 v_2), \nabla \cdot w_1) = 0, \quad \forall w_1 \in V_h^1,\\ (\kappa\nabla(v_2 - R_h^2 v_2), \nabla w_2) = 0, \quad \forall w_2 \in V_h^2.\end{aligned}$$

with error estimates (see [7, Lemma 2.2])

$$\|v_1 - R_h^1(v_1, v_2)\|_{H^1} \leq Ch\|D^2 v_1\| + C\|v_2 - R_h^2 v_2\|, \tag{3.15}$$

$$\|v_2 - R_h^2 v_2\|_{H^1} \leq Ch\|D^2 v_2\|. \tag{3.16}$$

The error can now be split according to

$$\begin{aligned}u_h^n - u(t_n) &= u_h^n - R_h^1(u(t_n), \theta(t_n)) + R_h^1(u(t_n), \theta(t_n)) - u(t_n) =: \eta_{h,u}^n + \rho_{h,u}^n,\\ \theta_h^n - \theta(t_n) &= \theta_h^n - R_h^2\theta(t_n) + R_h^2\theta(t_n) - \theta(t_n) =: \eta_{h,\theta}^n + \rho_{h,\theta}^n,\end{aligned}$$

where the error of $\rho_{h,u}^n$ and $\rho_{h,\theta}^n$ follows from (3.15)-(3.16). The first parts $\eta_{h,u}^n$ and $\eta_{h,\theta}^n$ can be plugged into the equation (2.4)-(2.5) to derive error estimates for these. Compare to the parabolic problem in Section 3.3. For the details we refer to [7].

To derive a GFEM for the thermoelasticity problem (1.8)-(1.9) we need to decompose two different spaces; V_h^1 and V_h^2. The decomposition of V_h^1 is performed with respect to the bilinear form $(\sigma(\cdot) : \varepsilon(\cdot))$ and the decomposition of V_h^2 with respect to $(\kappa\nabla\cdot, \nabla\cdot)$. This is done by mimicking the procedure described in Section 3.1. First define two interpolations $I_H^1 : V_h^1 \to V_H^1$ and $I_H^2 : V_h^2 \to V_H^2$ into the coarse finite element spaces $V_H^1 \subseteq V_h^1$ and $V_H^2 \subseteq V_h^2$. Now, the corresponding kernels are $V_{\mathrm{f}}^1 := \ker I_H^1$ and $V_{\mathrm{f}}^2 := \ker I_H^2$, and we can define the Ritz projections onto the these $R_{\mathrm{f}}^1 : V_h^1 \to V_{\mathrm{f}}^1$ and $R_{\mathrm{f}}^2 : V_h^2 \to V_{\mathrm{f}}^2$ given by

$$\begin{aligned}(\sigma(v_1 - R_{\mathrm{f}}^1 v_1) : \varepsilon(w_1)) = 0, \quad \forall w_1 \in V_{\mathrm{f}}^1,\ v_1 \in V_h^1\\ (\kappa\nabla(v_2 - R_{\mathrm{f}}^2 v_2), \nabla w_2) = 0, \quad \forall w_2 \in V_{\mathrm{f}}^2,\ v_2 \in V_h^2.\end{aligned}$$

The multiscale spaces are finally defined as

$$V_{\mathrm{ms}}^1 := V_H^1 - R_{\mathrm{f}}^1 V_H^1, \quad V_{\mathrm{ms}}^2 := V_H^2 - R_{\mathrm{f}}^2 V_H^2,$$

as in (3.5). With these spaces we can now define a Ritz projection corresponding to the stationary system. Define $R_{\mathrm{ms}}(v_1, v_2) : V_h^1 \times V_h^2 \to V_{\mathrm{ms}}^1 \times V_{\mathrm{ms}}^2$, such that,

$R_{\mathrm{ms}}(v_1,v_2)=(R^1_{\mathrm{ms}}(v_1,v_2),R^2_{\mathrm{ms}}v_2)$ and for all $(v_1,v_2)\in V^1_h\times V^2_h$,

$$\begin{aligned}(\sigma(v_1-R^1_{\mathrm{ms}}(v_1,v_2)):\varepsilon(w_1))-(\alpha(v_2-R^2_{\mathrm{ms}}v_2),\nabla\cdot w_1)&=0,\quad\forall w_1\in V^1_{\mathrm{ms}},\\(\kappa\nabla(v_2-R^2_{\mathrm{ms}}v_2),\nabla w_2)&=0,\quad\forall w_2\in V^2_{\mathrm{ms}}.\end{aligned}$$

The spaces V^1_{ms} and V^2_{ms} are designed to handle multiscale behavior in the coefficients μ,λ, and κ respectively. However, α is also material dependent and can be expected to vary at the same scale. For this reason, we shall add an extra correction to the solution $R_{\mathrm{ms}}(v_1,v_2)$ inspired by the techniques in [11, 8]. This additional correction is defined as $\tilde{R}_{\mathrm{f}}:V^2_h\to V^1_{\mathrm{f}}$, such that,

$$(\sigma(\tilde{R}_{\mathrm{f}}v_2):\varepsilon(w_1))=(\alpha R^2_{\mathrm{ms}}v_2,\nabla\cdot w_1),\quad\forall w_1\in V^1_{\mathrm{f}},$$

and we define $\tilde{R}^1_{\mathrm{ms}}(v_1,v_2)=R^1_{\mathrm{ms}}(v_1,v_2)+\tilde{R}_{\mathrm{f}}v_2$. Using the two operators $\mathcal{A}_1:V^1_h\times V^2_h\to V^1_h$ and $\mathcal{A}_2:V^2_h\to V^2_h$ defined by

$$\begin{aligned}(\mathcal{A}_1(v_1,v_2),w_1)&=(\sigma(v_1):\varepsilon(w_1))-(\alpha v_2,\nabla\cdot w_1),&&\forall w_1\in V^1_h,\\(\mathcal{A}_2v_2,w_2)&=(\kappa\nabla v_2,\nabla w_2),&&\forall w_2\in V^2_h,\end{aligned}$$

we prove, in Paper III, that the following error bounds hold for any $(v_1,v_2)\in V^1_h\times V^2_h$

$$\|v_1-\tilde{R}^1_{\mathrm{ms}}(v_1,v_2)\|_{H^1}\le CH\|\mathcal{A}_1(v_1,v_2)\|+C\|v_2-R^2_{\mathrm{ms}}v_2\|,\tag{3.17}$$

$$\|v_2-R^2_{\mathrm{ms}}v_2\|_{H^1}\le CH\|\mathcal{A}_2v_2\|,\tag{3.18}$$

where C is independent of the variations in μ,λ,α, and κ.

The following system defines a GFEM for the time dependent problem (2.4)-(2.5). For $n\in\{1,...,N\}$ find $\tilde{u}^n_{\mathrm{ms}}=u^n_{\mathrm{ms}}+u^n_{\mathrm{f}}$, with $u^n_{\mathrm{ms}}\in V^1_{\mathrm{ms}}$ and $u^n_{\mathrm{f}}\in V^1_{\mathrm{f}}$, and $\theta^n_{\mathrm{ms}}\in V^2_{\mathrm{ms}}$, such that

$$(\sigma(\tilde{u}^n_{\mathrm{ms}}):\varepsilon(v_1))-(\alpha\theta^n_{\mathrm{ms}},\nabla\cdot v_1)=(f^n,v_1),\quad\forall v_1\in V^1_{\mathrm{ms}},\tag{3.19}$$

$$(\bar{\partial}_t\theta^n_{\mathrm{ms}},v_2)+(\kappa\nabla\theta^n_{\mathrm{ms}},\nabla v_2)+(\alpha\nabla\cdot\bar{\partial}_t\tilde{u}^n_{\mathrm{ms}},v_2)=(g^n,v_2),\quad\forall v_2\in V^2_{\mathrm{ms}},\tag{3.20}$$

$$(\sigma(u^n_{\mathrm{f}}):\epsilon(w_1))-(\alpha\theta^n_{\mathrm{ms}},\nabla\cdot w_1)=0,\qquad\forall w_1\in V^1_{\mathrm{f}},\tag{3.21}$$

where $\tilde{u}^0_{\mathrm{ms}}=\tilde{u}_{\mathrm{ms},0}$ and $\theta^0_{\mathrm{ms}}=\theta_{\mathrm{ms},0}$ are suitable approximations of $u_{h,0}$ and $\theta_{h,0}$ (see Paper III). Here we have added an additional correction, u^n_{f}, on u^n_{ms} inspired by the correction in the stationary setting. Following the classical finite element analysis one can now split the error according to

$$\begin{aligned}\tilde{u}^n_{\mathrm{ms}}-u^n_h&=\tilde{u}^n_{\mathrm{ms}}-\tilde{R}^1_{\mathrm{ms}}(u^n_h,\theta^n_h)+\tilde{R}^1_{\mathrm{ms}}(u^n_h,\theta^n_h)-u^n_h=:\tilde{\eta}^n_{\mathrm{ms},u}+\rho^n_{\mathrm{ms},u},\\\theta^n_{\mathrm{ms}}-\theta^n_h&=\theta^n_{\mathrm{ms}}-R^2_{\mathrm{ms}}\theta^n_h+R^2_{\mathrm{ms}}\theta^n_h-\theta^n_h=:\eta^n_{\mathrm{ms},\theta}+\rho^n_{\mathrm{ms},\theta},\end{aligned}$$

where the error of $\tilde{\rho}^n_{\mathrm{ms},u}$ and $\rho^n_{\mathrm{ms},\theta}$ are bounded by (3.17)-(3.18). The error of $\tilde{\eta}^n_{u,\mathrm{ms}}$ and $\eta^n_{\theta,\mathrm{ms}}$ follows by plugging these into (3.19)-(3.21). However, in this case $\tilde{\eta}^n_{\mathrm{ms},u}\notin V^1_{\mathrm{ms}}$, which needs to be taken into account in the analysis.

To proceed we need to perform a localization of both spaces V^1_{ms} and V^2_{ms}. We use the patches $\omega_k(K)$ defined in Section 3.2 to define localized spaces $V^1_{\mathrm{ms},k}$ and $V^2_{\mathrm{ms},k}$, as in (3.8). To motivate this we need to show that the corrections $R^1_{\mathrm{f}}\lambda_x$ and $R^2_{\mathrm{f}}\lambda_y$ decay exponentially away from node x and y, where λ_x and λ_y

denotes the classical hat functions in V_H^1 and V_H^2 respectively. The correction $R_{\mathrm{f}}^2\lambda_y$ is based on the bilinear form $(\kappa\nabla\cdot, \nabla\cdot)$ of the same type as in Section 3.1 and the decay thus follows directly from [12, 8]. The correction $R_{\mathrm{f}}^1\lambda_x$ is based on the elasticity form $(\sigma(\cdot) : \varepsilon(\cdot))$ and the decay does *not* follow directly from the earlier results. This is instead proven in Paper II.

The localized GFEM for (2.4)-(2.5) is now defined as; for $n \in \{1, ..., N\}$ find

$$\tilde{u}_{\mathrm{ms},k}^n = u_{\mathrm{ms},k}^n + \sum_{K\in\mathcal{T}_H} u_{\mathrm{f},k}^{n,K}, \text{ with } u_{\mathrm{ms},k}^n \in V_{\mathrm{ms},k}^1,\ u_{\mathrm{f},k}^{n,K} \in V_{\mathrm{f}}^1(\omega_k(K)),$$

and $\theta_{\mathrm{ms},k}^n \in V_{\mathrm{ms},k}^2$, such that

$$(\sigma(\tilde{u}_{\mathrm{ms},k}^n) : \varepsilon(v_1)) - (\alpha\theta_{\mathrm{ms},k}^n, \nabla\cdot v_1) = (f^n, v_1), \quad \forall v_1 \in V_{\mathrm{ms},k}^1, \tag{3.22}$$

$$(\bar{\partial}_t\theta_{\mathrm{ms},k}^n, v_2) + (\kappa\nabla\theta_{\mathrm{ms},k}^n, \nabla v_2) + (\alpha\nabla\cdot\bar{\partial}_t\tilde{u}_{\mathrm{ms},k}^n, v_2) = (g^n, v_2), \quad \forall v_2 \in V_{\mathrm{ms},k}^2, \tag{3.23}$$

$$(\sigma(u_{\mathrm{f},k}^{n,K}) : \varepsilon(w_1)) - (\alpha\theta_{\mathrm{ms},k}^n, \nabla\cdot w_1)_K = 0, \qquad \forall w_1 \in V_{\mathrm{f}}^1(\omega_k(K)). \tag{3.24}$$

The main theorem in this thesis is Theorem 3.3 below and is proved in Paper III under certain conditions on the size of H. Here $C_{f,g}$ denotes a constant depending on f and g, see Paper III for details.

THEOREM 3.3. *Let $\{u_h^n\}_{n=1}^N$ and $\{\theta_h^n\}_{n=1}^N$ be the solutions to* (2.4)-(2.5) *and $\{\tilde{u}_{\mathrm{ms},k}^n\}_{n=1}^N$ and $\{\theta_{\mathrm{ms},k}^n\}_{n=1}^N$ the solutions to* (3.22)-(3.24). *For $n \in \{1, ..., N\}$ we have*

$$\|u_h^n - \tilde{u}_{\mathrm{ms},k}^n\|_{H^1} + \|\theta_h^n - \theta_{\mathrm{ms},k}^n\|_{H^1} \le C(H + k^{d/2}\xi^k)\big(C_{f,g} + t_n^{-1/2}\|\theta_h^0\|_{H^1}\big),$$

where C and $C_{f,g}$ are constants independent of the variations in σ, λ, α, and κ.

4. Summary of papers

Paper I. In Paper I we propose and analyze the GFEM (3.13) for parabolic equations with highly varying and oscillating coefficients. We prove convergence of optimal (second) order in the L_2-norm to the reference solution assuming initial data only in L_2. We do not assume any structural conditions on the multiscale coefficient, such as, periodicity or scale separation. Furthermore, we show how to extend this method to semilinear parabolic problems, where the right hand side in (3.10) is replaced by $f(u)$.

Paper II. In Paper II we propose a GFEM for linear elasticity equations with applications in heterogeneous materials. In particular, we prove exponential decay of the corrections $R_{\mathrm{f}}^1\lambda_z$ in Section 3.4. Furthermore, we prove that the GFEM reduces the locking effect that occur for materials with large Lamé parameter λ when using classical continuous and piecewise linear finite elements.

Paper III. In Paper III we build on the theory developed in Paper I and Paper II (originating from [12]) to define a GFEM for linear thermoelasticity with highly varying coefficients describing a heterogeneous material. We prove linear convergence to the reference solution in the H^1-norm independent of the variations in the data, see Theorem 3.3 in Section 3.4.

5. Future work

In Paper I on parabolic equations we assume that the diffusion coefficient $A(x)$ is independent of time. A natural extension would be to include time dependent coefficients $A(t,x)$. However, the main idea of the paper, to replace V_h with the space V_{ms} in (3.11), then fails. We would need to have a new space V_{ms}^n for each time t_n, since the diffusion coefficient $A(t_n,\cdot)$ takes different values for different times t_n. This is considerably more expensive than the time independent case, since we need to compute new corrections at each time step. It is possible that a more refined strategy could be developed by working with the parabolic problem in a space-time framework and perform localization in both time and space.

In applications involving composite materials there may be uncertainties in the material parameters, such as position or rotation, coming from the assembly procedure. These uncertainties can, for instance, be modeled by letting the coefficients depend on a random variable ω. A first step in extending the GFEM framework to such problems could be to consider an elliptic problem of the form

$$-\nabla \cdot A(x,\omega)\nabla u(x,\omega) = f(x,\omega),$$

where $A(\cdot,\omega)$ is multiscale in space for a fix ω. This problem suffers from the same problem as the time dependent case, since $A(\cdot,\omega)$ now takes different values for different outcomes ω.

In the analysis of the localization the constant $\xi \in (0,1)$, see e.g. Theorem 3.2, depends on the contrast β/α of A, that is, the ratio between the maximal and minimal value obtained by A. Also the constant C in Theorem 3.2 depends on this ratio. However, in available numerical examples, see Paper I and Paper II, but also, e.g., [12, 8], the size of the patches and the resulting convergence does not seem to be affected by large contrasts. Thus, the error bounds derived for the localization could be too crude. This should be further investigated to derive sharper error bounds for special classes of A.

References

[1] I. Babuška and R. Lipton: *Optimal local approximation spaces for generalized finite element methods with application to multiscale problems*, Multiscale Model. Simul. 9 (2011), no. 1, p. 373–406.

[2] I. Babuška and J. E. Osborn: *Generalized finite element methods: their performance and their relation to mixed methods*, SIAM J. Numer. Anal. 20 (1983), no. 3, 1983.

[3] M. A. Biot: *Thermoelasticity and irreversible thermodynamics*, J. Appl. Phys., 27 (1956), p. 240–253.

[4] B. A. Boley and J. H. Weiner: *Theory of thermal stresses*, John Wiley & Sons, Inc., New York London, 1960.

[5] E, Weinan and Engquist, Bjorn: *The heterogeneous multiscale methods*, Commun. Math. Sci. 1 (2003), no. 1, p. 87–132.

[6] Ch. Engwer and P. Henning and A. Målqvist and D. Peterseim: *Efficient implementation of the Localized Orthogonal Decomposition method*, Submitted.

[7] A. Ern and S. Meunier: *A posteriori error analysis of Euler-Galerkin approximations to coupled elliptic-parabolic problems*, M2AN Math. Model. Numer. Anal. 43 (2009), no. 2, p. 353 – 375.

[8] P. Henning and A. Målqvist: *Localized orthogonal decomposition techniques for boundary value problems*, SIAM J. Sci. Comput. 36 (2014), no. 4, p. A1609–A1634.

[9] T. Y. Hou and X.-H. Wu: *A Multiscale Finite Element Method for Elliptic Problems in Composite Materials and Porous Media*, J. Comput. Phys. 134 (1997), no. 1, p. 169–189.

[10] T. J. R. Hughes, G. R. Feijóo, L. Mazzei, and J-B. Quincy: *The variational multiscale method—a paradigm for computational mechanics*, Comput. Methods Appl. Mech. Engrg. 166 (1998), no. 1-2, p. 3–24.

[11] M. G. Larson and A. Målqvist: *Adaptive variational multiscale methods based on a posteriori error estimation: energy norm estimates for elliptic problems*, Comput. Methods Appl. Mech. Engrg. 196 (2007), no. 21-24, p. 2313–2324.

[12] A. Målqvist and D. Peterseim: *Localization of elliptic multiscale problems*, Math. Comp. 83 (2014), no. 290, p. 2583–2603.

[13] D. Peterseim: *Variational Multiscale Stabilization and the Exponential Decay of Fine-scale Correctors*, Submitted.

[14] R. E. Showalter: *Diffusion in poro-elastic media*, J. Math. Anal. Appl. 251 (2000), no. 1, p.310–340.

[15] V. Thomée: *Galerkin Finite Element Methods for Parabolic Problems*, Springer Series in Computational Mathematics, Springer-Verlag, Berlin, Second edition, 2009.

[16] A. Ženíšek: *Finite element methods for coupled thermoelasticity and coupled consolidation of clay*, RAIRO Anal. Numér. 18 (1984), no. 2, p. 183–205.

[17] A. Ženíšek: *The existence and uniqueness theorem in Biot's consolidation theory*, Apl. Mat. 29 (1984), no. 3, p. 194–211.

Part 2

Papers

Paper I

Axel Målqvist and Anna Persson, *Multiscale techniques for parabolic equations*, Preprint.

Multiscale techniques for parabolic equations

Axel Målqvist[1,2] and Anna Persson[1]

Abstract. We use the local orthogonal decomposition technique introduced in [15] to derive a generalized finite element method for linear and semilinear parabolic equations with spatial multiscale diffusion coefficient. We consider nonsmooth initial data and a backward Euler scheme for the temporal discretization. Optimal order convergence rate, depending only on the contrast, but not on the variations in the diffusion coefficient, is proven in the $L_\infty(L_2)$-norm. We present numerical examples, which confirm our theoretical findings.

1. Introduction

In this paper we study numerical solutions to a parabolic equation with a highly varying diffusion coefficient. These equations appear, for instance, when modeling physical behavior in a composite material or a porous medium. Such problems are often referred to as *multiscale problems*.

Convergence of optimal order of classical finite element methods (FEMs) based on continuous piecewise polynomials relies on at least spatial H^2 - regularity. More precisely, for piecewise linear polynomials, the error bound depends on $\|u\|_{H^2}$, where $\|u\|_{H^2} \sim \epsilon^{-1}$ if the diffusion coefficient varies on a scale of ϵ. Thus, the mesh width h must fulfill $h < \epsilon$ to achieve convergence. However, this is not computationally feasible in many applications. To overcome this issue, several numerical methods have been proposed, see, for example, [2], [5], [9], [15], [16], [17], and references therein. In particular, [16] and [17] consider linear parabolic equations.

In [15] a generalized finite element method (GFEM) was introduced and convergence of optimal order was proven for elliptic multiscale equations. The method builds on ideas from the variational multiscale method ([9],[11]), which is based on a decomposition of the solution space into a (coarse) finite dimensional space and a residual space for the fine scales. The method in [15], often referred to as local orthogonal decomposition, constructs a generalized finite element space where the basis functions contain information from the diffusion coefficient and have support on small vertex patches. With this approach, convergence of optimal order can be proved for an arbitrary positive and bounded diffusion coefficient. Restrictive assumptions such as periodicity of the coefficients or scale separation are not needed. Some recent works ([7], [8], [14]) show how this method can be applied to boundary value problems, eigenvalue problems, and semilinear elliptic equations. There has also been some recent work on the linear wave equation [1].

[1] Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg SE-412 96 Göteborg, Sweden.

[2] Supported by the Swedish Research Council.

In this paper we apply the technique introduced in [15] to parabolic equations with multiscale diffusion coefficients. For the discretization of the temporal domain we use the backward Euler scheme. Using tools from classical finite element theory for parabolic equations, see, e.g, [12], [13], [18], and references therein, we prove convergence of optimal order in the $L_\infty(L_2)$-norm for linear and semilinear equations under minimal regularity assumptions and nonsmooth initial data. The analysis is completed with numerical examples that support our theoretical findings.

In Section 2 we describe the problem formulation and the assumptions needed to achieve sufficient regularity of the solution. Section 3 describes the numerical approximation and presents the resulting GFEM. In Section 4 we prove error estimates and in Section 5 we extend the results to semilinear parabolic equations. Finally, in Section 6 we present some numerical examples.

2. Problem formulation

We consider the parabolic problem

$$
\begin{aligned}
\dot{u} - \nabla \cdot (A \nabla u) &= f, && \text{in } \Omega \times (0,T], \\
u &= 0, && \text{on } \partial\Omega \times (0,T], \\
u(\cdot,0) &= u_0, && \text{in } \Omega,
\end{aligned}
\tag{2.1}
$$

where $T > 0$ and Ω is a bounded polygonal/polyhedral domain in $\mathbb{R}^d$, $d \leq 3$. We assume $A = A(x)$ and $f = f(x,t)$, that is, the coefficient matrix A does not depend on the time variable.

We let $H^1(\Omega)$ denote the classical Sobolev space with norm

$$\|v\|^2_{H^1(\Omega)} = \|v\|^2_{L_2(\Omega)} + \|\nabla v\|^2_{L_2(\Omega)}$$

and $V = H^1_0(\Omega)$ the space of functions in $H^1(\Omega)$ that vanishes on $\partial\Omega$. We use $H^{-1}(\Omega) = V^*$ to denote the dual space to V. Furthermore, we use the notation $L_p(0,T;X)$ for the Bochner space with finite norm

$$
\begin{aligned}
\|v\|_{L_p(0,T;X)} &= \Big(\int_0^T \|v\|^p_X \,\mathrm{dt} \Big)^{1/p}, \quad 1 \leq p < \infty, \\
\|v\|_{L_\infty(0,T;X)} &= \operatorname*{ess\,sup}_{0 \leq t \leq T} \|v\|_X,
\end{aligned}
$$

where X is a Banach space equipped with norm $\|\cdot\|_X$. Here $v \in H^1(0,T;X)$ means $v, \dot{v} \in L_2(0,T;X)$. The dependence on the interval $[0,T]$ and the domain Ω is frequently suppressed and we write, for instance, $L_2(L_2)$ for $L_2(0,T;L_2(\Omega))$. Finally, we abbreviate the L_2-norm $\|\cdot\| := \|\cdot\|_{L_2(\Omega)}$ and the energy norm, $|||\cdot||| := \|A^{1/2}\nabla \cdot \|$.

To ensure existence, uniqueness, and sufficient regularity, we make the following assumptions on the data.

Assumptions. We assume

(A1) $A \in L^\infty(\Omega, \mathbb{R}^{d\times d})$, symmetric, and

$$0 < \alpha := \operatorname*{ess\,inf}_{x\in\Omega} \inf_{v\in\mathbb{R}^d\setminus\{0\}} \frac{A(x)v\cdot v}{v\cdot v},$$

$$\infty > \beta := \operatorname*{ess\,sup}_{x\in\Omega} \sup_{v\in\mathbb{R}^d\setminus\{0\}} \frac{A(x)v\cdot v}{v\cdot v},$$

(A2) $u_0 \in L_2$,

(A3) $f, \dot{f} \in L_\infty(L_2)$.

Throughout this work C denotes constants that may depend on the bounds α and β (often through the contrast β/α), the shape regularity parameter γ (3.1) of the mesh, the final time T, and the size of the domain Ω, but not on the mesh size parameters nor the derivatives of the coefficients in A. The fact that the constant does not depend on the derivatives of A is crucial, since these (if they exist) are large for the problems of interest. This is sometimes also noted as C being independent of *the variations of A*.

We now formulate the variational form of problem (2.1). Find $u(\cdot,t) \in V$ such that $u(\cdot,0) = u_0$ and

$$(\dot{u}, v) + a(u,v) = (f,v), \quad \forall v \in V,\ t \in (0,T], \tag{2.2}$$

where $(u,v) = \int_\Omega uv$ and $a(u,v) = (A\nabla u, \nabla v)$.

The following theorem states existence and uniqueness for (2.2). The proof is based on Galerkin approximations, see, e.g., [6] and [10].

Theorem 2.1. Assume that (A1), (A2), and (A3) holds. Then there exists a unique solution u to (2.2) such that $u \in L_2(0,T;V)$ and $\dot{u} \in L_2(0,T;H^{-1})$.

3. Numerical approximation

In this section we describe the local orthogonal decomposition presented in [15] to define a generalized finite element method for the multiscale problem (2.2).

First we introduce some notation. Let $\{\mathcal{T}_h\}_{h>0}$ and $\{\mathcal{T}_H\}_{H>h}$ be families of shape regular triangulations of Ω where $h_K := \operatorname{diam}(K)$, for $K \in \mathcal{T}_h$, and $H_K := \operatorname{diam}(K)$, for $K \in \mathcal{T}_H$. We also define $H := \max_{K\in\mathcal{T}_H} H_K$ and $h := \max_{K\in\mathcal{T}_h} h_K$. Furthermore, we let $\gamma > 0$ denote the shape regularity parameter of the mesh $\mathcal{T}_H$;

$$\gamma := \max_{K\in\mathcal{T}_H} \gamma_K, \text{ with } \gamma_K := \frac{\operatorname{diam} B_K}{\operatorname{diam} K}, \text{ for } K \in \mathcal{T}_H, \tag{3.1}$$

where B_K is the largest ball contained in K.

Now define the classical piecewise affine finite element spaces

$$V_H = \{v \in C(\bar{\Omega}) : v = 0 \text{ on } \partial\Omega, v|_K \text{ is a polynomial of degree} \le 1, \forall K \in \mathcal{T}_H\},$$

$$V_h = \{v \in C(\bar{\Omega}) : v = 0 \text{ on } \partial\Omega, v|_K \text{ is a polynomial of degree} \le 1, \forall K \in \mathcal{T}_h\}.$$

We let $\mathcal{N}$ denote the interior nodes of V_H and φ_x the corresponding hat function for $x \in \mathcal{N}$, such that $\operatorname{span}(\{\varphi_x\}_{x\in\mathcal{N}}) = V_H$. We further assume that $\mathcal{T}_h$ is a refinement of $\mathcal{T}_H$, such that $V_H \subseteq V_h$. Finally, we also need the finite element mesh $\mathcal{T}_H$ of Ω to be of a form such that the L_2-projection P_H onto the finite element space V_H is stable in H^1-norm, see, e.g., [3], and the references therein.

To discretize in time we introduce the uniform discretization

$$0 = t_0 < t_1 < ... < t_N = T, \text{ where } t_n - t_{n-1} = \tau. \tag{3.2}$$

Let U_n be the approximation of $u(t)$ at time $t = t_n$ and denote $f_n := f(t_n)$. Using the notation $\bar{\partial}_t U_n = (U_n - U_{n-1})/\tau$ we now formulate the classical backward Euler FEM; find $U_n \in V_h$ such that

$$(\bar{\partial}_t U_n, v) + a(U_n, v) = (f_n, v), \quad \forall v \in V_h, \tag{3.3}$$

for $n = 1, ..., N$ and $U_0 \in V_h$ is some approximation of u_0. For example, one could choose $U_0 = P_h u_0$, where P_h is the L_2-projection onto V_h. We also define the operator $\mathcal{A}_h : V_h \to V_h$ by

$$(\mathcal{A}_h v, w) = a(v, w), \quad \forall v, w \in V_h. \tag{3.4}$$

The convergence of the classical finite element approximation (3.3) depends on $\|D^2 u\|$, where D^2 denotes the second order derivatives. If the diffusion coefficient A oscillates on a scale of ϵ we have $\|D^2 u\| \sim \epsilon^{-1}$. Indeed, defining $\mathcal{A} = -\nabla \cdot A \nabla$, elliptic regularity gives

$$\begin{aligned}\|D^2 u\| &\leq C_1 \|\Delta u\| \leq C_2 \|A \Delta u\| \leq C_2 \|\nabla \cdot A \nabla u - \nabla A \cdot \nabla u\| \\ &\leq C_2(\|\mathcal{A} u\| + \|\nabla A \cdot \nabla u\|) \leq C_2(\|\mathcal{A} u\| + C_A \|\nabla u\|) \leq C_3(1 + C_A)\|\mathcal{A} u\|,\end{aligned}$$

where C_A is a constant that depends on the derivatives (variations) of A. This inequality is sharp in the sense that $\mathcal{A} u$ and $\nabla A \cdot \nabla u$ does not cancel in general. The total error is thus $\|u(t_n) - U_n\| \sim (\tau + (h/\epsilon)^2)$, which is small only if $h < \epsilon$.

The purpose of the method described in this paper is to find an approximate solution, let us denote it by $\hat{U}$ for now, in some space $\hat{V} \subset V_h$, such that $\dim \hat{V} = \dim V_H$, for $H > h$, and the error $\|U_n - \hat{U}_n\| \leq CH^2$. Here C is independent of the variations in A and $\hat{U}_n$ is less expensive to compute than U_n. The total error is then the sum of two terms

$$\|u(t_n) - \hat{U}_n\| \leq \|u(t_n) - U_n\| + \|U_n - \hat{U}_n\|,$$

where the first term is the error due to the standard FEM approximation with backward Euler discretization in time. This is small if h is chosen small enough, that is, if h resolves the variations of A. Hence, we think of $h > 0$ as fix and appropriately chosen. Our aim is now to analyze the error $\|U_n - \hat{U}_n\|$.

We emphasize that $\hat{V} = V_H$ is not sufficient. The total error would in this case be $\|u(t_n) - \hat{U}_n\| \sim (\tau + (H/\epsilon)^2)$, which is small only if $H < \epsilon$.

The next theorem states some regularity results for (3.3).

Theorem 3.1. Assume that (A1), (A2), and (A3) holds. Then, for $1 \leq n \leq N$, there exists a unique solution U_n to (2.2) such that $U_n \in V_h$. Furthermore, if $U_0 = 0$, then we have the bound

$$\|\bar{\partial}_t U_n\| \leq C(\|f\|_{L_\infty(L_2)} + \|\dot{f}\|_{L_\infty(L_2)}), \tag{3.5}$$

and, if $f = 0$, then

$$\|\bar\partial_t U_n\| \le C t_n^{-1}\|U_0\|,\ n \ge 1, \quad \|\bar\partial_t\bar\partial_t U_n\| \le C t_n^{-2}\|U_0\|,\ n \ge 2, \tag{3.6}$$

where C depends on α and T, but not on the variations of A.

Proof. From (3.3) it follows for $n \ge 2$ that

$$(\bar\partial_t\bar\partial_t U_n, v) + a(\bar\partial_t U_n, v) = (\bar\partial_t f_n, v), \quad \forall v \in V_h,$$

and the stability estimate for backward Euler schemes gives

$$\|\bar\partial_t U_n\| \le \|\bar\partial_t U_1\| + \sum_{j=2}^{n} \tau\|\bar\partial_t f_j\|.$$

From (3.3) we have, since $U_0 = 0$, $\|\bar\partial_t U_1\| \le \|f_1\|$. Finally, using the inequality

$$\sum_{j=2}^{n} \tau\|\bar\partial_t f_j\| \le \sum_{j=2}^{n} \max_{t_{j-1}\le\xi\le t_j} \tau\|\dot f(\xi)\| \le C\|\dot f\|_{L_\infty(L_2)},$$

we deduce (3.5).

For the bound (3.6) we refer to [18, Lemma 7.3]. □

3.1. Orthogonal decomposition. In this section we describe the orthogonal decomposition which defines the GFEM space denoted $\hat V$ in the discussion above. We refer to [15] and [14] for details.

For the construction of the GFEM space we use the (weighted) Clément interpolation operator introduced in [4], $\mathfrak{I}_H \colon V_h \to V_H$ defined by

$$\mathfrak{I}_H v = \sum_{x\in\mathcal{N}} (\mathfrak{I}_H v)(x)\varphi_x, \quad \text{where} \quad (\mathfrak{I}_H v)(x) := \frac{\int_\Omega v\varphi_x}{\int_\Omega \varphi_x}. \tag{3.7}$$

For this interpolation operator the following result is proved [4]

$$H_K^{-1}\|v - \mathfrak{I}_H v\|_{L_2(K)} + \|\nabla(v - \mathfrak{I}_H v)\|_{L_2(K)} \le C\|\nabla v\|_{L_2(\bar\omega_K)}, \forall v \in V, \tag{3.8}$$

where $\bar\omega_K := \cup\{\bar K \in \mathcal{T}_H : \bar K \cap K \ne \emptyset\}$ and C depends on the shape regularity γ.

Let $V^{\mathrm{f}} = \{v \in V_h : \mathfrak{I}_H v = 0\}$ be the kernel of the Clément interpolation operator (3.7). This space contains all fine scale features not resolved by V_H. The space V_h can then be decomposed into $V_h = V_H \oplus V^{\mathrm{f}}$, where $v \in V_h$ can be written as a sum $v = v_H + v^{\mathrm{f}}$, with $v_H \in V_H$, $v^{\mathrm{f}} \in V^{\mathrm{f}}$, and $(v_H, v^{\mathrm{f}}) = 0$.

Now define the orthogonal projection $R^{\mathrm{f}} \colon V_h \to V^{\mathrm{f}}$ by

$$a(R^{\mathrm{f}} v, w) = a(v, w) \quad \forall w \in V^{\mathrm{f}},\ v \in V_h.$$

Using this projection we define the GFEM space, also referred to as the multiscale space,

$$V^{\mathrm{ms}} := V_H - R^{\mathrm{f}} V_H,$$

which leads to another orthogonal decomposition $V_h = V^{\mathrm{ms}} \oplus V^{\mathrm{f}}$. Hence any function $v \in V_h$ has a unique decomposition $v = v^{\mathrm{ms}} + v^{\mathrm{f}}$, with $v^{\mathrm{ms}} \in V^{\mathrm{ms}}$ and $v^{\mathrm{f}} \in V^{\mathrm{f}}$, with $a(v^{\mathrm{ms}}, v^{\mathrm{f}}) = 0$.

To define a basis for V^{ms} we need to find the projection R^{f} of the nodal basis function $\varphi_x \in V_H$. Let this projection be denoted ϕ_x, so that $\phi_x \in V^{\mathrm{f}}$ satisfies the (global) corrector problem

$$a(\phi_x, w) = a(\varphi_x, w), \quad \forall w \in V^{\mathrm{f}}. \tag{3.9}$$

A basis for the multiscale space V^{ms} is thus given by

$$\{\varphi_x - \phi_x \colon x \in \mathcal{N}\}.$$

We also introduce the projection $R^{\mathrm{ms}} \colon V_h \to V^{\mathrm{ms}}$, defined by

$$a(R^{\mathrm{ms}}v, w) = a(v, w), \quad \forall w \in V^{\mathrm{ms}},\ v \in V_h. \tag{3.10}$$

Note that $R^{\mathrm{ms}} = I - R^{\mathrm{f}}$. For R^{ms} we have the following lemma, based on the results in [15].

Lemma 3.2. For the projection R^{ms} in (3.10) and $v \in V_h$ we have the error bound

$$\|v - R^{\mathrm{ms}}v\| \le CH^2\|\mathcal{A}_h v\|, \quad v \in V_h, \tag{3.11}$$

where C depends on α and γ, but not on the variations of A.

Proof. Define the following elliptic auxiliary problem: find $z \in V_h$ such that

$$a(z, w) = (v - R^{\mathrm{ms}}v, w), \quad \forall w \in V_h.$$

In [15, Lemma 3.1] it was proven that the solution to an elliptic equation of the form

$$a(u, w) = (g, w), \quad \forall w \in V_h,$$

satisfies the error estimate

$$|||u - R^{\mathrm{ms}}u||| \le CH\|g\|,$$

where C depends on γ and α, but not on the variations of A. Hence, we have the following bound for z,

$$|||z - R^{\mathrm{ms}}z||| \le CH\|v - R^{\mathrm{ms}}v\|.$$

Furthermore, we note that $v - R^{\mathrm{ms}}v \in V_h$ and

$$\begin{aligned}\|v - R^{\mathrm{ms}}v\|^2 &= (v - R^{\mathrm{ms}}v, v - R^{\mathrm{ms}}v) = a(z, v - R^{\mathrm{ms}}v)\\ &= a(z - R^{\mathrm{ms}}z, v - R^{\mathrm{ms}}v) \le |||z - R^{\mathrm{ms}}z|||\ |||v - R^{\mathrm{ms}}v|||.\end{aligned}$$

Now, since $a(v, w) = (\mathcal{A}_h v, w)$, we get $|||v - R^{\mathrm{ms}}v||| \le CH\|\mathcal{A}_h v\|$ and (3.11) follows. □

In particular, if U_n is the solution to (3.3), then (3.11) gives

$$\|U_n - R^{\mathrm{ms}}U_n\| \le CH^2\|P_h f_n - \bar\partial_t U_n\|, \quad n \ge 1,$$
$$\|\bar\partial_t U_n - R^{\mathrm{ms}}\bar\partial_t U_n\| \le CH^2\|P_h\bar\partial_t f_n - \bar\partial_t\bar\partial_t U_n\|, \quad n \ge 2.$$

The result in Lemma 3.2 should be compared with the error of the classical Ritz projection $R_h : V \to V_h$ defined by $a(R_h v, w) = a(v, w)$, $\forall w \in V_h$. Using elliptic regularity estimates, one achieves

$$\|R_h v - v\| \leq Ch^2 \|D^2 v\| \leq Ch^2 \|\mathcal{A} v\|,$$

which is similar to the result in Lemma 3.2. However, in this case, C depends on the variations of A, as we noted in the discussion in the beginning of this section. This is avoided using the R^{ms}-projection, since the constant in Lemma 3.2 does not depend on the variations of A.

Now let P^{ms} denote the L_2-projection onto V^{ms} and define the corresponding GFEM to problem (3.3); find $U_n^{\mathrm{ms}} \in V^{\mathrm{ms}}$ such that $U_0^{\mathrm{ms}} = P^{\mathrm{ms}} U_0$ and

$$(\bar{\partial}_t U_n^{\mathrm{ms}}, v) + a(U_n^{\mathrm{ms}}, v) = (f_n, v), \quad \forall v \in V^{\mathrm{ms}}, \tag{3.12}$$

for $n = 1, ..., N$. Furthermore, we define the operator $\mathcal{A}^{\mathrm{ms}} : V^{\mathrm{ms}} \to V^{\mathrm{ms}}$ by

$$(\mathcal{A}^{\mathrm{ms}} v, w) = a(v, w), \quad \forall v, w \in V^{\mathrm{ms}}. \tag{3.13}$$

3.2. Localization. Since the corrector problems (3.9) are posed in the fine scale space V^{f} they are computationally expensive to solve. Moreover, the correctors ϕ_x generally have global support, which destroys the sparsity of the resulting linear system (3.12). However, as shown in [15], ϕ_x decays exponentially fast away from x. This observation motivates a localization of the corrector problems to smaller patches of coarse elements. Here we use a variant presented in [7], which reduces the required size of the patches.

We first define the notion of patches and their sizes. For all $K \in \mathcal{T}_H$ we define $\omega_k(K)$ to be the patch of size k, where

$$\begin{aligned}
&\omega_0(K) := K, \\
&\omega_k(K) := \cup\{\bar{K} \in \mathcal{T}_H : \bar{K} \cap \omega_{k-1}(K) \neq \emptyset\}, \quad k = 1, 2, ...
\end{aligned}$$

Moreover, we define $V^{\mathrm{f}}(\omega_k(K)) := \{w \in V^{\mathrm{f}} : \mathrm{supp}(w) \subset \omega_k(K)\}$.

Now define the operator $R_K^{\mathrm{f}} \colon V_h \to V^{\mathrm{f}}$ by

$$\int_\Omega A \nabla R_K^{\mathrm{f}} v \cdot \nabla w = \int_K A \nabla v \cdot \nabla w, \quad \forall v \in V_h, \ w \in V^{\mathrm{f}},$$

and note that $R^{\mathrm{f}} := \sum_{K \in \mathcal{T}_H} R_K^{\mathrm{f}}$. We now localize the operator R_K^{f} by defining $R_{K,k}^{\mathrm{f}} \colon V_h \to V^{\mathrm{f}}(\omega_k(K))$ through

$$\int_{\omega_k(K)} A \nabla R_{K,k}^{\mathrm{f}} v \cdot \nabla w = \int_K A \nabla v \cdot \nabla w, \quad \forall v \in V_h, \ w \in V^{\mathrm{f}}(\omega_k(K)),$$

and we define $R_k^{\mathrm{f}} := \sum_{K \in \mathcal{T}_H} R_{K,k}^{\mathrm{f}}$. Hence we can, for each nonnegative integer k, define a localized multiscale space

$$V_k^{\mathrm{ms}} := V_H - R_k^{\mathrm{f}} V_H.$$

Here the basis is given by $\{\varphi_x - \phi_{k,x} : x \in \mathcal{N}\}$, where $\phi_{k,x} = R_k^{\mathrm{f}} \varphi_x$ is the localized version of ϕ_x. The procedure of decomposing V_h into the orthogonal spaces V^{ms} and V^{f} together with the localization of V^{ms} to V_k^{ms} is referred to as *local orthogonal decomposition.*

The following lemma follows from Lemma 3.6 in [7].

Lemma 3.3. There exists a constant $0 < \mu < 1$ that depends on the contrast β/α such that

$$|||R^{\mathrm{f}}v - R_k^{\mathrm{f}}v||| \leq Ck^{d/2}\mu^k |||v|||, \quad \forall v \in V_h,$$

where C depends on β, α, and γ, but not on the variations of A.

Now let $R_k^{\mathrm{ms}}\colon V_h \to V_k^{\mathrm{ms}}$ be the orthogonal projection defined by

$$a(R_k^{\mathrm{ms}}v, w) = a(v, w), \quad \forall w \in V_k^{\mathrm{ms}}. \tag{3.14}$$

Next lemma is a consequence of Theorem 3.7 in [7] and estimates the error due to the localization procedure.

Lemma 3.4. For the projection R_k^{ms} in (3.14) we have the bound

$$\|v - R_k^{\mathrm{ms}}v\| \leq C(H + k^{d/2}\mu^k)^2 \|\mathcal{A}_h v\|, \quad \forall v \in V_h. \tag{3.15}$$

Here C depends on β, α, and γ, but not on the variations of A.

Proof. The proof is similar to the proof of Lemma 3.2. Let $z \in V_h$ be the solution to the elliptic dual problem

$$a(z, w) = (v - R_k^{\mathrm{ms}}v, w), \quad \forall w \in V_h,$$

which gives

$$\begin{aligned}\|v - R_k^{\mathrm{ms}}v\|^2 &= (v - R_k^{\mathrm{ms}}v, v - R_k^{\mathrm{ms}}v) = a(z - R_k^{\mathrm{ms}}z, v - R_k^{\mathrm{ms}}v)\\ &\leq |||z - R_k^{\mathrm{ms}}z||| |||v - R_k^{\mathrm{ms}}v|||.\end{aligned}$$

It follows from Theorem 3.7 in [7] that there exists a constant C depending on β, α, and γ, such that $|||z - R_k^{\mathrm{ms}}z||| \leq C(H + k^{d/2}\mu^k)\|v - R_k^{\mathrm{ms}}v\|$, with μ as in Lemma 3.3. Since $(\mathcal{A}_h v, w) = a(v, w)$ we get $|||v - R_k^{\mathrm{ms}}v||| \leq C(H + k^{d/2}\mu^k)\|\mathcal{A}_h v\|$ and (3.15) follows. □

We are now ready to formulate the localized version of (3.12) by replacing V^{ms} by V_k^{ms}. The localized GFEM formulation reads; find $U_{k,n}^{\mathrm{ms}} \in V_k^{\mathrm{ms}}$ such that $U_{k,0}^{\mathrm{ms}} = P_k^{\mathrm{ms}} U_0$ and

$$(\bar{\partial}_t U_{k,n}^{\mathrm{ms}}, v) + a(U_{k,n}^{\mathrm{ms}}, v) = (f_n, v), \quad \forall v \in V_k^{\mathrm{ms}}, \tag{3.16}$$

for $n = 1, ..., N$, where P_k^{ms} is the L_2-projection onto V_k^{ms}. We also define the operator $\mathcal{A}_k^{\mathrm{ms}} : V_k^{\mathrm{ms}} \to V_k^{\mathrm{ms}}$, a localized version of (3.13), by

$$(\mathcal{A}_k^{\mathrm{ms}} v, w) = a(v, w), \quad \forall v, w \in V_k^{\mathrm{ms}}. \tag{3.17}$$

Next lemma states some important properties of the operators $\mathcal{A}_h$, $\mathcal{A}^{\mathrm{ms}}$, and $\mathcal{A}_k^{\mathrm{ms}}$.

Lemma 3.5. The operators $\mathcal{A}_h$, $\mathcal{A}^{\rm ms}$, and $\mathcal{A}_k^{\rm ms}$ are self-adjoint and positive definite. Furthermore, the following bound hold

$$\|(\mathcal{A}_k^{\rm ms})^{-1/2} P_k^{\rm ms} f\| \leq C\|f\|_{H^{-1}}, \quad \forall f \in L_2,$$

where C depends on α.

Proof. The fact that the operators are self-adjoint and positive definite follows from the assumptions on A (A1). A proof of the bound can be found in [13]. □

We also define the solution operator $E_{k,n}^{\rm ms} = ((I+\tau\mathcal{A}_k^{\rm ms})^{-1})^n$, such that the solution to (3.16), with $f=0$, can be expressed as $U_{k,n}^{\rm ms} = E_{k,n}^{\rm ms} U_{k,0}^{\rm ms}$. For this operator we have estimates similar to (3.6).

Lemma 3.6. For $l = 0, 1$, and $v \in L_2$, we have

$$\|\bar{\partial}_t^l E_{k,n}^{\rm ms} P_k^{\rm ms} v\| \leq C t_n^{-l} \|v\|, \quad n \geq l, \quad |||E_{k,n}^{\rm ms} P_k^{\rm ms} v||| \leq C t_n^{-1/2} \|v\|, \quad n \geq 1,$$

where C depends on the constant $C_k = \sup_{s>0} s^k e^{-s}$, β, and α.

Proof. The operator $\mathcal{A}_k^{\rm ms}$ is self-adjoint, positive definite, and defined on the finite dimensional space $V_k^{\rm ms}$. Thus, there exist a finite number of positive eigenvalues $\{\lambda_i\}_{i=1}^M$ and corresponding orthogonal eigenvectors $\{\varphi_i\}_{i=1}^M$ such that $\text{span}\{\varphi_i\} = V_k^{\rm ms}$. We refer to [14] for a further discussion on the eigenvalues to the operator $A_k^{\rm ms}$.

It follows that $E_{k,n}^{\rm ms} v$ can be written as

$$E_{k,n}^{\rm ms} v = \sum_{i=1}^M \frac{1}{(1+\tau\lambda_i)^n} (v, \varphi_i)\varphi_i,$$

and the estimates now follows from [18, Lemma 7.3]. □

4. Error analysis

In this section we derive error estimates for the local orthogonal decomposition method introduced in Section 3. The localized GFEM solution (3.16) is compared to the classical FEM solution (3.3), which leads to a setting where the initial data is not smooth, since $U_0 \in V_h$ only. This leads to error bounds which are non-uniform in time, but of optimal order for a fix time $t_n > 0$. The same phenomenon appears in classical finite element analysis for equations with nonsmooth initial data, see [18] and references therein. The error analysis in this section is carried out by only taking the L_2-norm of U_0, which allows $u_0 \in L_2$. If we, for instance, choose $U_0 = P_h u_0$, then $\|U_0\| \leq \|u_0\|$.

Theorem 4.1. Let U_n be the solution to (3.3) and $U_{k,n}^{\rm ms}$ the solution to (3.16). Then, for $1 \leq n \leq N$,

$$\|U_{k,n}^{\rm ms} - U_n\| \leq C\Big(1 + \log\frac{t_n}{\tau}\Big)(H + k^{d/2}\mu^k)^2\big(t_n^{-1}\|U_0\| + \|f\|_{L_\infty(L_2)} + \|\dot{f}\|_{L_\infty(L_2)}\big),$$

where C depends on β, α, γ, and T, but not on the variations of A.

The proof of Theorem 4.1 is divided into two lemmas. The first covers the homogeneous case, $f = 0$, and the second covers the nonhomogeneous case with vanishing initial data $u_0 = 0$. To study the error in the homogeneous case we use techniques similar to the classical finite element analysis of problems with nonsmooth initial data, see [18] and the references therein.

Define $T_h = \mathcal{A}_h^{-1} P_h$ and $T_k^{\mathrm{ms}} = (\mathcal{A}_k^{\mathrm{ms}})^{-1} P_k^{\mathrm{ms}}$. With this notation the solution to the parabolic problem (3.3), with $f = 0$, can be expressed as $T_h \bar{\partial}_t U_n + U_n = 0$. Similarly, the solution to (3.16), with $f = 0$, can be expressed as $T_k^{\mathrm{ms}} \bar{\partial}_t U_{k,n}^{\mathrm{ms}} + U_{k,n}^{\mathrm{ms}} = 0$. Note that T_k^{ms} is self-adjoint and positive semi-definite on L_2, and that $T_k^{\mathrm{ms}} = R_k^{\mathrm{ms}} T_h$.

Now, let $e_n = U_{k,n}^{\mathrm{ms}} - U_n$, where e_n solves the error equation

$$T_k^{\mathrm{ms}} \bar{\partial}_t e_n + e_n = -U_n - T_k^{\mathrm{ms}} \bar{\partial}_t U_n = (T_h - T_k^{\mathrm{ms}}) \bar{\partial}_t U_n = (R_k^{\mathrm{ms}} - I) U_n =: \rho_n, \tag{4.1}$$

for $n = 1, ..., N$ with $T_k^{\mathrm{ms}} e_0 = 0$, since $U_{k,0}^{\mathrm{ms}} = P_k^{\mathrm{ms}} U_0$. The following lemma is a discrete versions of [18, Lemma 3.3].

Lemma 4.2. Suppose e_n satisfies the error equation (4.1). Then

$$\|e_n\|^2 \leq C\Big(\|\rho_n\|^2 + t_n^{-1}\Big(\sum_{j=1}^{n} \tau \|\rho_j\|^2 + \sum_{j=2}^{n} \tau t_j^2 \|\bar{\partial}\rho_j\|^2\Big)\Big), \quad n \geq 2, \tag{4.2}$$

$$\|e_1\| \leq \|\rho_1\|. \tag{4.3}$$

Proof. Multiply the error equation (4.1) by $\bar{\partial}_t e_n$ and integrate over Ω to get

$$(T_k^{\mathrm{ms}} \bar{\partial}_t e_n, \bar{\partial}_t e_n) + (e_n, \bar{\partial}_t e_n) = (\rho_n, \bar{\partial}_t e_n),$$

where the first term on the left hand side is nonnegative, since T_k^{ms} is positive semi-definite on L_2. Multiplying by τt_n we have

$$t_n \|e_n\|^2 - t_n(e_n, e_{n-1}) \leq t_n(\rho_n, e_n - e_{n-1}),$$

which gives

$$\begin{aligned}\frac{t_n}{2}\|e_n\|^2 - \frac{t_{n-1}}{2}\|e_{n-1}\|^2 &\leq t_n(\rho_n, e_n - e_{n-1}) + \frac{t_n - t_{n-1}}{2}\|e_{n-1}\|^2 \\ &\leq t_n(\rho_n, e_n) - t_{n-1}(\rho_{n-1}, e_{n-1}) \\ &\quad - (t_n\rho_n - t_{n-1}\rho_{n-1}, e_{n-1}) + \frac{\tau}{2}\|e_{n-1}\|^2.\end{aligned}$$

Summing over n now gives

$$\begin{aligned}t_n\|e_n\|^2 - t_1\|e_1\|^2 \leq 2t_n(\rho_n, e_n) - 2t_1(\rho_1, e_1) - \sum_{j=2}^{n} 2(t_j\rho_j - t_{j-1}\rho_{j-1}, e_{j-1}) \\ + \sum_{j=2}^{n} \tau \|e_{j-1}\|^2,\end{aligned}$$

and thus,

$$t_n\|e_n\|^2 \le C\Big(t_n\|\rho_n\|^2 + \sum_{j=2}^{n} \tau\big(t_j^2\|\bar{\partial}_t\rho_j\|^2 + \|\rho_{j-1}\|^2\big) + \sum_{j=2}^{n}\tau\|e_{j-1}\|^2\Big).$$

To estimate the last sum we note that, since T_k^{ms} is self-adjoint and positive semi-definite,

$$\begin{aligned}2(T_k^{\mathrm{ms}}\bar{\partial}_t e_n, e_n) &= (T_k^{\mathrm{ms}}\bar{\partial}_t e_n, e_n) + (T_k^{\mathrm{ms}} e_n, \bar{\partial}_t e_n)\\ &= \bar{\partial}_t(T_k^{\mathrm{ms}} e_n, e_n) + \tau(T_k^{\mathrm{ms}}\bar{\partial}_t e_n, \bar{\partial}_t e_n) \ge \bar{\partial}_t(T_k^{\mathrm{ms}} e_n, e_n).\end{aligned}$$

so by multiplying the error equation (4.1) by $2e_n$ we get

$$\bar{\partial}_t(T_k^{\mathrm{ms}} e_n, e_n) + 2\|e_n\|^2 \le 2(T_k^{\mathrm{ms}}\bar{\partial}_t e_n, e_n) + 2\|e_n\|^2 = 2(\rho_n, e_n).$$

Multiplying by τ and summing over n gives

$$(T_k^{\mathrm{ms}} e_n, e_n) + \sum_{j=1}^{n}\tau\|e_j\|^2 \le \sum_{j=1}^{n}\tau\|\rho_j\|^2,$$

where we have used that $T_k^{\mathrm{ms}} e_0 = 0$. Since the first term is nonnegative we deduce that $\sum_{j=1}^{n}\tau\|e_j\|^2 \le \sum_{j=1}^{n}\tau\|\rho_j\|^2$ and (4.2) follows. For $n = 1$ this also proves (4.3). □

Next Lemma is a discrete version of a result that can be found in the proof of [18, Theorem 3.3].

LEMMA 4.3. Under the assumptions of Lemma 4.2 we have, for $n \ge 2$, the bound

$$\|e_n\| \le C t_n^{-1}\Big(\max_{2\le j\le n} t_j^2\|\bar{\partial}_t\rho_j\| + \max_{1\le j\le n}\Big(t_j\|\rho_j\| + \|\sum_{r=1}^{j}\tau\rho_r\|\Big)\Big). \tag{4.4}$$

PROOF. It follows from Lemma 4.2 that

$$\|e_n\| \le C(\max_{2\le j\le n} t_j\|\bar{\partial}_t\rho_j\| + \max_{1\le j\le n}\|\rho_j\|), \quad n \ge 2,$$

or by using Young's inequality with different constants the proof can be modified to show that

$$\|e_n\| \le \epsilon\max_{2\le j\le n} t_j\|\bar{\partial}_t\rho_j\| + C(\epsilon)\max_{1\le j\le n}\|\rho_j\|, \quad n \ge 2,$$

for some $\epsilon > 0$. Now define $z_j = t_j e_j$. Then

$$T_k^{\mathrm{ms}}\bar{\partial}_t z_n + z_n = t_n\rho_n + T_k^{\mathrm{ms}} e_{n-1} := \eta_n, \quad n \ge 1,$$

and, since $T_k^{\mathrm{ms}} z_0 = 0$ we conclude from Lemma 4.2

$$\|z_n\| \le \epsilon\max_{2\le j\le n} t_j\|\bar{\partial}_t\eta_j\| + C\max_{1\le j\le n}\|\eta_j\|.$$

From the definition of η_j it follows that

$$\|\eta_j\| \le t_j\|\rho_j\| + \|T_k^{\mathrm{ms}} e_{j-1}\|, \quad j \ge 1.$$

Furthermore, for $j \geq 2$

$$\begin{aligned} t_j \|\bar{\partial}_t \eta_j\| &\leq t_j \|\bar{\partial}_t t_j \rho_j\| + t_j \|\bar{\partial}_t T_k^{\mathrm{ms}} e_{j-1}\| \\ &\leq t_j^2 \|\bar{\partial}_t \rho_j\| + t_j \|\rho_{j-1}\| + t_j \|\rho_{j-1} - e_{j-1}\| \\ &\leq t_j^2 \|\bar{\partial}_t \rho_j\| + 2t_j \|\rho_j - \rho_{j-1}\| + 2t_j \|\rho_j\| + t_j \|e_{j-1}\| \\ &\leq 3t_j^2 \|\bar{\partial}_t \rho_j\| + 2t_j \|\rho_j\| + 2t_{j-1} \|e_{j-1}\| \\ &\leq C\Big(t_j^2 \|\bar{\partial}_t \rho_j\| + t_j \|\rho_j\|\Big) + 2\|z_{j-1}\|, \end{aligned}$$

where we used $\frac{1}{2} t_j \leq t_{j-1} \leq t_j$ for $j \geq 2$. To bound $\|T_k^{\mathrm{ms}} e_n\|$ we define $\tilde{e}_n = \sum_{j=1}^n \tau e_j$ and $\tilde{e}_0 = 0$. Multiplying the error equation (4.1) by τ and summing over n gives

$$\sum_{j=1}^n \tau T_k^{\mathrm{ms}} \bar{\partial}_t e_n + \tilde{e}_n = T_k^{\mathrm{ms}} \bar{\partial}_t \tilde{e}_n + \tilde{e}_n = \tilde{\rho}_n, \quad n \geq 1,$$

where $\tilde{\rho}_n = \sum_{j=1}^n \tau \rho_j$ and we have used that $T_k^{\mathrm{ms}} e_0 = 0$. Note that by definition $T_k^{\mathrm{ms}} \tilde{e}_0 = 0$. Thus, by Lemma 4.2, we have

$$\begin{aligned} \|\tilde{e}_n\| &\leq C\Big(\max_{2 \leq j \leq n} t_j \|\bar{\partial}_t \tilde{\rho}_j\| + \max_{1 \leq j \leq n} \|\tilde{\rho}_j\|\Big) \\ &\qquad\qquad \leq C\Big(\max_{2 \leq j \leq n} t_j \|\rho_j\| + \max_{1 \leq j \leq n} \|\sum_{r=1}^j \tau \rho_r\|\Big). \end{aligned}$$

Hence, since $T_k^{\mathrm{ms}} \bar{\partial}_t \tilde{e}_n = T_k^{\mathrm{ms}} e_n$,

$$\|T_k^{\mathrm{ms}} e_n\| \leq \|\tilde{e}_n\| + \|\tilde{\rho}_n\| \leq C\Big(\max_{2 \leq j \leq n} t_j \|\rho_j\| + \max_{1 \leq j \leq n} \|\sum_{r=1}^j \tau \rho_r\|\Big).$$

With $\epsilon = \frac{1}{4}$ we get

$$\begin{aligned} \|z_n\| &\leq \frac{1}{4} \max_{2 \leq j \leq n} t_j \|\bar{\partial}_t \eta_j\| + C \max_{1 \leq j \leq n} \|\eta_j\| \\ &\leq \frac{1}{2} \max_{1 \leq j \leq n} \|z_j\| + C\Big(\max_{2 \leq j \leq n} t_j^2 \|\bar{\partial}_t \rho_j\| + \max_{1 \leq j \leq n} (t_j \|\rho_j\| + \|\sum_{r=1}^j \tau \rho_r\|)\Big), \end{aligned}$$

but from (4.3) we deduce $\|z_1\| \leq t_1 \|\rho_1\|$, and hence

$$\|z_n\| \leq \frac{1}{2} \max_{2 \leq j \leq n} \|z_j\| + C\Big(\max_{2 \leq j \leq n} t_j^2 \|\bar{\partial}_t \rho_j\| + \max_{1 \leq j \leq n} (t_j \|\rho_j\| + \|\sum_{r=1}^j \tau \rho_r\|)\Big).$$

Choosing n^* such that $\max_{2 \leq j \leq n} z_j = z_{n^*}$ we conclude (4.4). □

Lemma 4.4. Assume $f = 0$ and let $U_{k,n}^{\mathrm{ms}}$ be the solution to (3.16) and U_n the solution to (3.3). Then, for $1 \leq n \leq N$,

$$\|U_{k,n}^{\mathrm{ms}} - U_n\| \leq C(H + k^{d/2} \mu^k)^2 t_n^{-1} \|U_0\|$$

where C depends on β, α, γ, and T, but not on the variations of A.

Proof. From Lemma 4.3 we have

$$\|e_n\| \leq Ct_n^{-1}\Big(\max_{2\leq j\leq n} t_j^2\|\bar{\partial}_t\rho_j\| + \max_{1\leq j\leq n}\Big(t_j\|\rho_j\| + \|\sum_{r=1}^{j}\tau\rho_r\|\Big)\Big), \quad n\geq 2,$$

and from Lemma 4.2 $\|e_1\| \leq \|\rho_1\|$. The rest of the proof is based on estimates for the projection R_k^{ms} in Lemma 3.4 and the regularity of the homogeneous equation (3.6). We have

$$\begin{aligned}
t_j^2\|\bar{\partial}_t\rho_j\| &\leq C(H+k^{d/2}\mu^k)^2 t_j^2\|\mathcal{A}_h\bar{\partial}_t U_j\| \\
&\leq C(H+k^{d/2}\mu^k)^2 t_j^2\|\bar{\partial}_t\bar{\partial}_t U_j\| \leq C(H+k^{d/2}\mu^k)^2\|U_0\|, \quad j\geq 2, \\
t_j\|\rho_j\| &\leq C(H+k^{d/2}\mu^k)^2 t_j\|\mathcal{A}_h U_j\| \leq C(H+k^{d/2}\mu^k)^2\|U_0\|, \quad j\geq 1, \\
\|\sum_{r=1}^{j}\tau\rho_r\| &= \|\sum_{r=1}^{j}\tau(T_h - T_k^{\mathrm{ms}})\bar{\partial}_t U_r\| \leq \|(T_h - T_k^{\mathrm{ms}})(U_j - U_0)\| \\
&\leq C(H+k^{d/2}\mu^k)^2\|U_0\|,
\end{aligned}$$

where we have used $\|U_j\| \leq \|U_0\|$, which completes the proof. □

The next lemma concerns the convergence of the inhomogeneous parabolic problem (2.1) with initial data $U_0 = 0$.

Lemma 4.5. Assume $U_0 = 0$ and let $U_{k,n}^{\mathrm{ms}}$ be the solution to (3.16) and U_n the solution to (3.3). Then, for $1 \leq n \leq N$,

$$\|U_{k,n}^{\mathrm{ms}} - U_n\| \leq C(1+\log\frac{t_n}{\tau})(H+k^{d/2}\mu^k)^2(\|f\|_{L_\infty(L_2)} + \|\dot{f}\|_{L_\infty(L_2)}),$$

where C depends on β, α, γ, and T, but not on the variations of A.

Proof. Let $U_{k,n}^{\mathrm{ms}} - U_n = U_{k,n}^{\mathrm{ms}} - R_k^{\mathrm{ms}}U_n + R_k^{\mathrm{ms}}U_n - U_n =: \theta_n + \rho_n$. For ρ_n we use Lemma 3.4 to achieve the estimate

$$\|\rho_n\| \leq C(H+k^{d/2}\mu^k)^2\|\mathcal{A}_h U_n\|.$$

Now, for $v \in V_k^{\mathrm{ms}}$ we have

$$\big(\bar{\partial}_t\theta_n, v\big) + a(\theta_n, v) = (-\bar{\partial}_t\rho_n, v).$$

Using Duhamel's principle we have

$$\theta_n = \tau\sum_{j=1}^{n} E_{k,n-j+1}^{\mathrm{ms}} P_k^{\mathrm{ms}}(-\bar{\partial}_t\rho_j),$$

since $\theta_0 = 0$. Summation by parts now gives

$$\theta_n = E_{k,n}^{\mathrm{ms}} P_k^{\mathrm{ms}}\rho_0 - P_k^{\mathrm{ms}}\rho_n + \tau\sum_{j=1}^{n}\bar{\partial}_t E_{k,n-j+1}^{\mathrm{ms}} P_k^{\mathrm{ms}}\rho_j.$$

Note that $\rho_0 = 0$. Using Lemma 3.4 and Lemma 3.6 we get

$$\begin{aligned}\|\theta_n\| &\leq \|\rho_n\| + \tau \sum_{j=1}^{n} t_{n-j+1}^{-1} \|\rho_j\| \\ &\leq C(H + k^{d/2}\mu^k)^2 \max_{1\leq j\leq n} \|\mathcal{A}_h U_j\| (1 + \tau \sum_{j=1}^{n} t_{n-j+1}^{-1}),\end{aligned}$$

where the last sum can be bounded by

$$\tau \sum_{j=1}^{n} t_{n-j+1}^{-1} \leq 1 + \log \frac{t_n}{\tau}.$$

It remains to bound $\|\mathcal{A}_h U_n\|$. We have $\mathcal{A}_h U_n = P_h f_n - \bar{\partial}_t U_n$ and Lemma 3.1 gives

$$\|\mathcal{A}_h U_j\| \leq \|f_j\| + \|\bar{\partial}_t U_j\| \leq C(\|f\|_{L_\infty(L_2)} + \|\dot{f}\|_{L_\infty(L_2)}),$$

which completes the proof. □

Proof of Theorem 4.1. The result follows from Lemma 4.4 and Lemma 4.5 by rewriting $U_n = U_{n,1} + U_{n,2}$, where $U_{n,1}$ is the solution to the homogeneous problem and $U_{n,2}$ the solution to the inhomogeneous problem with vanishing initial data. □

Remark 4.6. We note that the choice of k and the size of μ determine the rate of the convergence. In general, to achieve optimal order convergence rate, k should be chosen proportional to $\log(H^{-1})$, i.e. $k = c\log(H^{-1})$. With this choice of k we have $\|U_{k,n}^{\mathrm{ms}} - U_n\| \leq C(1 + \log n) H^2 t_n^{-1}$.

5. The semilinear parabolic equation

In this section we discuss how the above techniques can be extended to a semilinear parabolic problem with multiscale diffusion coefficient.

5.1. Problem formulation. We are interested in equations of the form

$$\begin{aligned} \dot{u} - \nabla \cdot (A \nabla u) &= f(u), && \text{in } \Omega \times (0,T], \\ u &= 0, && \text{on } \partial\Omega \times (0,T], \\ u(\cdot, 0) &= u_0, && \text{in } \Omega, \end{aligned} \tag{5.1}$$

where $f : \mathbb{R} \to \mathbb{R}$ is twice continuously differentiable and Ω is a polygonal/polyhedral boundary in $\mathbb{R}^d$, for $d \leq 3$. For $d = 2, 3$, f is assumed to fulfill the growth condition

$$|f^{(l)}(\xi)| \leq C(1 + |\xi|^{\delta+1-l}), \quad \text{for } l = 1, 2, \tag{5.2}$$

where $\delta = 2$ if $d = 3$ and $\delta \in [1, \infty)$ if $d = 2$. Furthermore, we assume that the diffusion A fulfills assumption (A1) and $u_0 \in V$.

Example 5.1. The Allen-Cahn equation $\dot{u} - \nabla \cdot (A \nabla u) = -(u^3 - u)$ fulfills the assumption (5.2).

Define the ball $B_R := \{v \in V : \|v\|_{H^1} \leq R\}$. Using Hölder and Sobolev inequalities the following lemma can be proved, see [13].

LEMMA 5.2. If f fulfills assumption (5.2) and $u, v \in B_R$, then

$$\begin{aligned} \|f(u)\| &\leq C, & \|f'(u)z\|_{H^{-1}} &\leq C\|z\|, \\ \|f'(u)z\| &\leq C\|z\|_{H^1}, & \|f''(u)z\|_{H^{-1}} &\leq C\|z\|, \end{aligned}$$

and

$$\|f(u) - f(v)\|_{H^{-1}} \leq C\|u - v\|,$$

where C is a constant depending on R.

From (5.1) we derive the variational form; find $u(t) \in V$ such that

$$(\dot{u}, v) + (A\nabla u, \nabla v) = (f(u), v), \quad \forall v \in V, \tag{5.3}$$

and $u(0) = u_0$. For this problem local existence of a solution can be derived given that the initial data $u_0 \in V$, see [13].

THEOREM 5.3. Assume that (A1) and (5.2) holds. Then, for $u_0 \in B_R$, there exist $\tau_0 = \tau_0(R)$ and $c > 0$, such that (5.3) has a unique solution $u \in C(0, \tau_0; V)$ and $\|u\|_{L_\infty(0,\tau_0;V)} \leq cR$.

For the Allen-Cahn equation it is possible to find an a priori global bound of u. This means that for any time T there exists R such that if u is a solution then $\|u(t)\|_{L_\infty(H^1)} \leq R$ for $t \in [0, T]$. Thus we can apply the local existence theorem repeatedly to attain global existence, see [13].

5.2. Numerical approximation.

The assumptions and definitions of the families of triangulations $\{\mathcal{T}_h\}_{h>0}$ and $\{\mathcal{T}_H\}_{H>h}$ and the corresponding spaces V_H and V_h remain the same as in Section 3. For the discretization in time we use a uniform time discretization given by

$$0 = t_0 < t_1 < ... < t_N = \tau_0, \text{ where } t_n - t_{n-1} = \tau, \tag{5.4}$$

where τ_0 is given from Theorem 5.3. With these discrete spaces we consider the semi-implicit backward Euler scheme where $U_n \in V_h$ satisfies

$$(\bar{\partial}_t U_n, v) + (A\nabla U_n, \nabla v) = (f(U_{n-1}), v), \quad \forall v \in V_h, \tag{5.5}$$

for $n = 1, ..., N$ where $U_0 \in V_h$ is an approximation of u_0. It is proven in [12] that this scheme satisfies the bound

$$\|U_n - u(t_n)\| \leq Ct_n^{-1/2}(h^2 + \tau),$$

if we choose, for instance, $U_0 = P_h u_0$, where P_h denotes the L_2-projection onto V_h. Note that C in this bound depends on the variations of A.

The following theorem gives some regularity estimates of the solution to (5.5).

Theorem 5.4. Assume that (A1) and (5.2) holds. Then, for $U_0 \in B_R$, there exist $\tau_0 = \tau_0(R)$ and $c > 0$ such that (5.5) has a unique solution $U_n \in V_h$, for $1 \le n \le N$, and $\max_{1\le n\le N} \|U_n\|_{H^1} \le cR$. Moreover, the following bounds hold

$$\begin{aligned} \|\bar\partial_t U_n\| &\le Ct_n^{-1/2},\ n\ge 1, \quad |||\bar\partial_t U_n||| \le Ct_n^{-1},\ n\ge 1,\\ \|\bar\partial_t\bar\partial_t U_n\| &\le Ct_n^{-3/2},\ n\ge 2, \end{aligned}$$

where C depends on α, τ_0, and R, but not on the variations of A.

Proof. We only prove the estimate $\|\bar\partial_t\bar\partial_t U_n\| \le Ct_n^{-3/2}$ here. The other two follow by similar arguments.

From (5.5) we get

$$(\bar\partial_t\bar\partial_t U_n, v) + a(\bar\partial_t U_n, v) = (\bar\partial_t f(U_{n-1}), v), \quad \forall v\in V_h,\ n\ge 2, \tag{5.6}$$

$$(\bar\partial_t^{(3)} U_n, v) + a(\bar\partial_t\bar\partial_t U_n, v) = (\bar\partial_t\bar\partial_t f(U_{n-1}), v), \quad \forall v\in V_h,\ n\ge 3. \tag{5.7}$$

Choosing $v = \bar\partial_t\bar\partial_t U_n$ in (5.7) gives

$$\frac{1}{\tau}\|\bar\partial_t\bar\partial_t U_n\|^2 - \frac{1}{\tau}(\bar\partial_t\bar\partial_t U_{n-1}, \bar\partial_t\bar\partial_t U_n) + |||\bar\partial_t\bar\partial_t U_n|||^2 = (\bar\partial_t\bar\partial_t f(U_{n-1}), \bar\partial_t\bar\partial_t U_n),$$

which gives the bound

$$\|\bar\partial_t\bar\partial_t U_n\|^2 - \|\bar\partial_t\bar\partial_t U_{n-1}\|^2 \le C\tau\|\bar\partial_t\bar\partial_t f(U_{n-1})\|_{H^{-1}}. \tag{5.8}$$

Using Lemma 5.2 we have for

$$\xi_j \in (\min\{U_{n-j}, U_{n-(j-1)}\}, \max\{U_{n-j}, U_{n-(j-1)}\})$$

the following bound

$$\begin{aligned} \|\bar\partial_t\bar\partial_t f(U_n)\|_{H^{-1}} &= \frac{1}{\tau^2}\|f'(\xi_1)(U_n - U_{n-1}) - f'(\xi_2)(U_{n-1} - U_{n-2})\|_{H^{-1}}\\ &\le \frac{1}{\tau^2}\|(f'(\xi_1) - f'(\xi_2))(U_n - U_{n-1})\|_{H^{-1}}\\ &\quad + \frac{1}{\tau^2}\|f'(\xi_2)(U_n - 2U_{n-1} + U_{n-2})\|_{H^{-1}}\\ &\le \frac{1}{\tau^2}\|(\xi_1 - \xi_2)(U_n - U_{n-1})\| + C\|\bar\partial_t\bar\partial_t U_n\|, \end{aligned}$$

Note that $|\xi_1 - \xi_2\| \le \|U_{n-2} - U_{n-1}\| + \|U_{n-1} - U_n\|$. By using Sobolev embeddings we get

$$\begin{aligned} \frac{1}{\tau^2}\|(\xi_1 - \xi_2)(U_n - U_{n-1})\| &\le \max_{n-1\le j\le n} 2\|(\bar\partial_t U_j)^2\| \le \max_{n-1\le j\le n} 2\|\bar\partial_t U_j\|_{L_4}^2\\ &\le C\max_{n-1\le j\le n}\|\bar\partial_t U_j\|_{H^1}^2 \le Ct_{n-1}^{-2} \le Ct_n^{-2}, \end{aligned}$$

where we recall the bounds $\frac{1}{2}t_j \leq t_{j-1} \leq t_j$ for $j \geq 2$. Multiplying by τt_n^4 in (5.8) and summing over n gives

$$\begin{aligned}
&t_n^4\|\bar{\partial}_t\bar{\partial}_t U_n\|^2 \\
&\quad\leq t_2^4\|\bar{\partial}_t\bar{\partial}_t U_2\|^2 + \sum_{j=3}^{n}(\tau t_j^4\|\bar{\partial}_t\bar{\partial}_t f(U_{n-1})\|_{H^{-1}}^2 + (t_j^4 - t_{j-1}^4)\|\bar{\partial}_t\bar{\partial}_t U_{j-1}\|^2) \\
&\quad\leq t_2^4\|\bar{\partial}_t\bar{\partial}_t U_2\|^2 + C\sum_{j=3}^{n}\tau\big(t_j^4\|\bar{\partial}_t\bar{\partial}_t U_{j-1}\|^2 + t_j^4 t_{j-1}^{-4} + t_{j-1}^3\|\bar{\partial}_t\bar{\partial}_t U_{j-1}\|^2\big) \\
&\quad\leq t_2^4\|\bar{\partial}_t\bar{\partial}_t U_2\|^2 + Ct_n + C\sum_{j=3}^{n}\tau\big(t_{j-1}^4\|\bar{\partial}_t\bar{\partial}_t U_{j-1}\|^2 + t_{j-1}^3\|\bar{\partial}_t\bar{\partial}_t U_{j-1}\|^2\big),
\end{aligned}$$

for $n \geq 3$. Using $\|\bar{\partial}_t U_j\| \leq Ct_j^{-1/2}$ for $j \geq 1$ we get

$$t_2^4\|\bar{\partial}_t\bar{\partial}_t U_2\|^2 \leq C\tau^2(\|\bar{\partial}_t U_2\|^2 + \|\bar{\partial}_t U_1\|^2) \leq C\tau^2(t_2^{-1} + t_1^{-1}) \leq C\tau.$$

Now, to bound $\sum_{j=2}^{n} t_j^3\|\bar{\partial}_t\bar{\partial}_t U_j\|$, we choose $v = \bar{\partial}_t\bar{\partial}_t U_n$ in (5.6) to derive

$$\|\bar{\partial}_t\bar{\partial}_t U_n\|^2 + \frac{1}{\tau}|||\bar{\partial}_t U_n|||^2 - \frac{1}{\tau}|||\bar{\partial}_t U_{n-1}|||^2 \leq \|\bar{\partial}_t f(U_{n-1})\|^2. \tag{5.9}$$

and with ξ_j as above, we get

$$\|\bar{\partial}_t f(U_{n-1})\| = \|f'(\xi_2)\bar{\partial}_t U_{n-1}\| \leq C|||\bar{\partial}_t U_{n-1}||| \leq Ct_{n-1}^{-1},$$

where we used Lemma 5.2 and $|||\bar{\partial}_t U_j||| \leq Ct_j^{-1}$ for $j \geq 1$. Multiplying 5.9 with τt_n^3 and summing over n gives

$$\begin{aligned}
\sum_{j=2}^{n}\tau t_j^3\|\bar{\partial}_t\bar{\partial}_t U_j\|^2 + t_n^3|||\bar{\partial}_t U_n|||^2 &\leq C\sum_{j=2}^{n}(\tau t_j^3 t_{j-1}^{-2} + (t_j^3 - t_{j-1}^3)|||\bar{\partial}_t U_{j-1}|||^2) \\
&\quad + t_1^3|||\bar{\partial}_t U_1|||^2 \\
&\leq C\sum_{j=2}^{n}(\tau t_j + \tau t_{j-1}^2|||\bar{\partial}_t U_{j-1}|||^2) + t_1^3|||\bar{\partial}_t U_1|||^2.
\end{aligned}$$

Using $|||\bar{\partial}_t U_j||| \leq Ct_j^{-1}$ for $j \geq 1$ we get

$$\sum_{j=2}^{n}\tau t_j^3\|\bar{\partial}_t\bar{\partial}_t U_j\|^2 \leq C(t_n^2 + t_n + t_1) \leq Ct_n,$$

where C now depends on $t_n \leq T$. So we have proved

$$\begin{aligned} t_n^4 \|\bar{\partial}_t \bar{\partial}_t U_n\|^2 &\leq C \sum_{j=3}^{n} \tau t_{j-1}^4 \|\bar{\partial}_t \bar{\partial}_t U_{j-1}\|^2 + C t_n + \tau \\ &\leq C \sum_{j=2}^{n-1} \tau t_j^4 \|\bar{\partial}_t \bar{\partial}_t U_j\|^2 + C t_{n+1} \\ &\leq C \sum_{j=2}^{n-1} \tau t_j^4 \|\bar{\partial}_t \bar{\partial}_t U_j\|^2 + C t_n. \end{aligned}$$

Applying the classical discrete Grönwall's lemma gives

$$t_n^4 \|\bar{\partial}_t \bar{\partial}_t U_n\|^2 \leq C t_n,$$

which proves $\|\bar{\partial}_t \bar{\partial}_t U_n\| \leq C t_n^{-3/2}$ for $n \geq 3$. For $n = 2$ we proved

$$t_2^4 \|\bar{\partial}_t \bar{\partial}_t U_2\|^2 \leq C\tau \leq C t_2,$$

which completes the proof. □

We use the same GFEM space as in Section 3, that is, $V^{\mathrm{ms}} = V_H - R^{\mathrm{f}}(V_H)$ and the localized version $V_k^{\mathrm{ms}} = V_H - R_k^{\mathrm{f}}(V_H)$. Furthermore, for the completely discrete scheme, we consider the time discretization defined in (5.4) and the linearized backward Euler method thus reads; find $U_{k,n}^{\mathrm{ms}} \in V^{\mathrm{ms}}$ such that $U_{k,0}^{\mathrm{ms}} = P_k^{\mathrm{ms}} U_0$ and

$$(\bar{\partial}_t U_{k,n}^{\mathrm{ms}}, v) + a(U_{k,n}^{\mathrm{ms}}, v) = (f(U_{k,n-1}^{\mathrm{ms}}), v), \tag{5.10}$$

for $n = 1, ..., N$ where P_k^{ms} is the L_2-projection onto V_k^{ms}.

To derive an error estimates we represent the solution to (5.10) by using Duhamel's principle. Note that $U_{k,n}^{\mathrm{ms}}$ is the solution to the equation

$$\bar{\partial}_t U_{k,n}^{\mathrm{ms}} + \mathcal{A}_k^{\mathrm{ms}} U_{k,n}^{\mathrm{ms}} = P_k^{\mathrm{ms}} f(U_{k,n-1}^{\mathrm{ms}}),$$

and by Duhamel's principle we get

$$U_{k,n}^{\mathrm{ms}} = E_{k,n}^{\mathrm{ms}} U_{k,0}^{\mathrm{ms}} + \tau \sum_{j=1}^{n} E_{k,n-j+1}^{\mathrm{ms}} P_k^{\mathrm{ms}} f(U_{k,j-1}^{\mathrm{ms}}).$$

5.3. Error analysis. For the error analysis we need the following generalized discrete Grönwall lemma, see, e.g., [13].

Lemma 5.5. Let $A, B \geq 0$, $\gamma_1, \gamma_2 > 0$, $0 \leq t_0 < t_n \leq T$, and $0 \leq \varphi_n \leq R$. If

$$\varphi_n \leq A t_n^{-1+\gamma_1} + B\tau \sum_{j=1}^{n-1} t_{n-j+1}^{-1+\gamma_2} \varphi_j,$$

then there is a constant C depending on B, γ_1, γ_2, and, T, such that,

$$\varphi_n \leq A t_n^{-1+\gamma_1}.$$

Theorem 5.6. For given $R \geq 0$ and $\tau_0 > 0$ let U_n be the solution to (5.5) and $U_{k,n}^{\mathrm{ms}}$ be the solution to (5.10), such that $U_n, U_{k,n}^{\mathrm{ms}} \in B_R$. Then, for $1 \leq n \leq N$,

$$\|U_{k,n}^{\mathrm{ms}} - U_n\| \leq C(H + k^{d/2}\mu^k)^2 t_n^{-1/2}, \tag{5.11}$$

where C depends on β, α, γ, R, and τ_0, but not on the variations of A.

Proof. First we define $e_n = U_{k,n}^{\mathrm{ms}} - U_n = (U_{k,n}^{\mathrm{ms}} - R_k^{\mathrm{ms}} U_n) + (R_k^{\mathrm{ms}} U_n - U_n) = \theta_n + \rho_n$. For ρ_j we use Lemma 3.4 to prove the bounds

$$\|\rho_j\| \leq C(H + k^{d/2}\mu^k)^2 t_j^{-1/2}, \quad j \geq 1,$$

and

$$\|\bar{\partial}_t \rho_j\| \leq C(H + k^{d/2}\mu^k)^2 t_j^{-3/2}, \quad j \geq 2.$$

For θ_n we have

$$\theta_n = E_{k,n}^{\mathrm{ms}}\theta_0 + \tau \sum_{j=1}^{n} E_{k,n-j+1}^{\mathrm{ms}} P_k^{\mathrm{ms}}(f(U_{k,j-1}^{\mathrm{ms}}) - f(U_{j-1}) - \bar{\partial}_t \rho_j).$$

To bound $\|\theta_n\|$ we first assume $n \geq 2$ and use summation by parts for the first part of the sum. Defining n_2 to be the integer part of $n/2$ we can write

$$\begin{aligned} -\tau \sum_{j=1}^{n_2} E_{k,n-j+1}^{\mathrm{ms}} P_k^{\mathrm{ms}} \bar{\partial}_t \rho_j &= E_{k,n}^{\mathrm{ms}} P_k^{\mathrm{ms}} \rho_0 - E_{k,n-n_2}^{\mathrm{ms}} P_k^{\mathrm{ms}} \rho_{n_2} \\ &\quad + \tau \sum_{j=1}^{n_2} \big(\bar{\partial}_t E_{k,n-j+1}^{\mathrm{ms}}\big) P_k^{\mathrm{ms}} \rho_j, \end{aligned}$$

and θ_n can be rewritten as

$$\begin{aligned} \theta_n = E_{k,n}^{\mathrm{ms}} P_k^{\mathrm{ms}} e_0 - E_{k,n-n_2}^{\mathrm{ms}} P_k^{\mathrm{ms}} \rho_{n_2} + \tau \sum_{j=1}^{n_2} \big(\bar{\partial}_t E_{k,n-j+1}^{\mathrm{ms}}\big) P_k^{\mathrm{ms}} \rho_j \\ - \tau \sum_{j=n_2+1}^{n} E_{k,n-j+1}^{\mathrm{ms}} P_k^{\mathrm{ms}} \bar{\partial}_t \rho_j \\ + \tau \sum_{j=1}^{n} (\mathcal{A}_k^{\mathrm{ms}})^{1/2} E_{k,n-j+1}^{\mathrm{ms}} (\mathcal{A}_k^{\mathrm{ms}})^{-1/2} P_k^{\mathrm{ms}} (f(U_{k,j-1}^{\mathrm{ms}}) - f(U_{j-1})), \end{aligned}$$

where we note that $P_k^{\mathrm{ms}} e_0 = 0$. To estimate these terms we need the following bounds for $\gamma_1, \gamma_2 > 0$

$$\tau \sum_{j=1}^{n} t_{n-j+1}^{-1+\gamma_1} t_j^{-1+\gamma_2} \leq C_{\gamma_1,\gamma_2} t_n^{-1+\gamma_1+\gamma_2}, \quad \tau \sum_{j=1}^{n_2} t_{n-j+1}^{-\gamma_1} t_j^{-1+\gamma_2} \leq C_{\gamma_1,\gamma_2} t_n^{-\gamma_1+\gamma_2}.$$

see [12]. Using Lemma 3.6 we get

$$\|\theta_n\| \leq \|\rho_{n_2}\| + C\tau \sum_{j=1}^{n_2} t_{n-j+1}^{-1} \|\rho_j\| + C\tau \sum_{j=n_2+1}^{n} \|\bar{\partial}_t \rho_j\| \\ + C\tau \sum_{j=1}^{n} t_{n-j+1}^{-1/2} \|f(U_{k,j-1}^{\mathrm{ms}}) - f(U_{j-1})\|_{H^{-1}},$$

and together with Lemma 3.4 and Lemma 5.2 this gives

$$\begin{aligned} \|\theta_n\| &\leq C(H + k^{d/2}\mu^k)^2 \Big(t_{n_2}^{-1/2} + \tau \sum_{j=1}^{n_2} t_{n-j+1}^{-1} t_j^{-1/2} + \tau \sum_{j=n_2+1}^{n} t_j^{-3/2} \Big) \\ &\quad + C\tau \sum_{j=1}^{n} t_{n-j+1}^{-1/2} \|U_{k,j-1}^{\mathrm{ms}} - U_{j-1}\| \\ &\leq C(H + k^{d/2}\mu^k)^2 t_n^{-1/2} + C\tau \sum_{j=1}^{n} t_{n-j+1}^{-1/2} \|e_{j-1}\|. \end{aligned}$$

Now consider θ_1. We can rewrite

$$\begin{aligned} \theta_1 &= E_{k,1}^{\mathrm{ms}} \theta_0 + \tau E_{k,1}^{\mathrm{ms}} P_k^{\mathrm{ms}} (f(U_{k,0}^{\mathrm{ms}}) - f(U_0) - \bar{\partial}_t \rho_1) \\ &= E_{k,1}^{\mathrm{ms}} P_k^{\mathrm{ms}} e_0 - E_{k,1}^{\mathrm{ms}} P_k^{\mathrm{ms}} \rho_1 + \tau E_{k,1}^{\mathrm{ms}} P_k^{\mathrm{ms}} (f(U_{k,0}^{\mathrm{ms}}) - f(U_0)), \end{aligned}$$

and using similar arguments as above

$$\|\theta_1\| \leq C(H + k^{d/2}\mu^k)^2 t_1^{-1/2} + \tau t_1^{-1/2} \|e_0\|,$$

Hence, we arrive at the estimate

$$\|e_n\| \leq C t_n^{-1/2} (H + k^{d/2}\mu)^2 + C\tau \sum_{j=1}^{n} t_{n-j+1}^{-1/2} \|e_{j-1}\|, \quad n \geq 1,$$

and we can use Lemma 5.5 to conclude (5.11). □

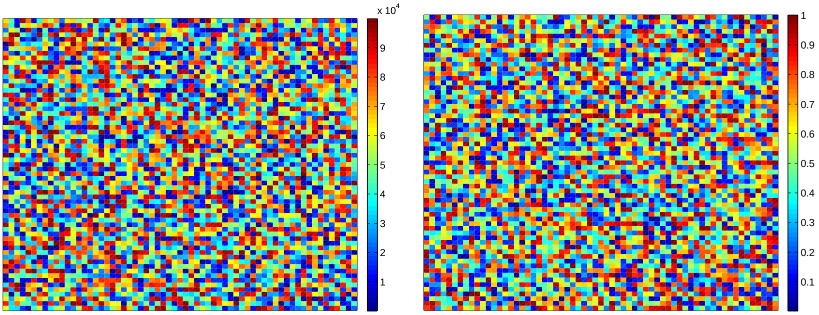
6. Numerical Results

In this section we present some numerical results to verify the predicted error estimates presented for the linear problem in Section 4 and the semilinear problem in Section 5. In both cases the domain is set to the unit square $\Omega = [0,1] \times [0,1]$ and $T = 1$. The domain Ω is discretized with a uniform triangulation and the interval $[0,T]$ is divided into subintervals of equal length.

The method is tested on two different types of diffusion coefficients A_1 and A_2 defined as

$$A_1(x) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad A_2(x) = \begin{pmatrix} B(x) & 0 \\ 0 & B(x) \end{pmatrix},$$

where B is piecewise constant with respect to a uniform Cartesian grid of size 2^{-6}, see Figure 1. Note that our choice of B imposes significant multiscale behavior on the diffusion coefficient. Here we expect quadratic convergence in space of the standard finite element with piecewise linear and continuous polynomials (P1-FEM) when $A = A_1$, but poor convergence when $A = A_2$. For the GFEM we expect quadratic convergence in both cases.



(A) Coefficient B for the linear parabolic problem. The contrast is $\beta/\alpha \approx 10^6$.

(B) Coefficient B for the semilinear parabolic problem. The contrast is $\beta/\alpha \approx 10^3$.

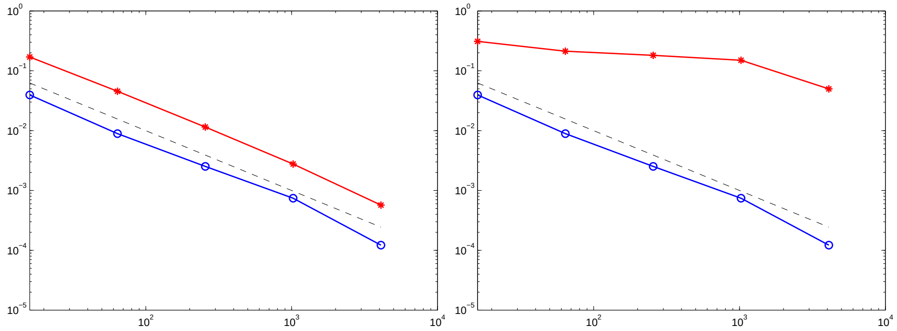
FIGURE 1. Coefficients for the two parabolic problems.

We compute the localized GFEM in (3.16) and (5.10), denoted $U^{\mathrm{ms}}_{k,n}$, for 5 different values of the coarse grid width, $H = \sqrt{2}\cdot 2^{-2}, \sqrt{2}\cdot 2^{-3}, \sqrt{2}\cdot 2^{-4}, \sqrt{2}\cdot 2^{-5}$, and $\sqrt{2}\cdot 2^{-6}$. The time step is chosen to $\tau = 0.01$ for all problems. The reference mesh $\mathcal{T}_h$ is of size $h = \sqrt{2}\cdot 2^{-7}$ and defines the space V_h on which the localized corrector problems $\phi_{k,x}$ are solved. To measure the error, the solution U_n in (3.3) is computed using P1-FEM on the finest scale $h = \sqrt{2}\cdot 2^{-7}$ with $\tau = 0.01$.

Note that this experiment measures the error $\|U_n - U^{\mathrm{ms}}_{k,n}\|$. The total error $\|u(t_n) - U^{\mathrm{ms}}_{k,n}\|$ is also affected by the difference $\|u(t_n) - U_n\|$, which is dominating for the smaller values of H. We now present the result in two separate sections.

6.1. Linear parabolic problem. For the linear parabolic problem (2.1) the right hand side is set to $f(x,t) = t$, which fulfills the assumptions for the required regularity. For simplicity the initial data is set to $u_0 = 1$. Moreover, at each cell in the Cartesian grid we choose a value from the interval $[10^{-1}, 10^5]$. This procedure gives B a rapidly varying feature and a high contrast $\max(B)/\min(B) \approx 10^6$, see Figure 1 (left).

For each value of H the localized GFEM, $U^{\mathrm{ms}}_{k,n}$, and the corresponding P1-FEM, denoted $U_{H,n}$, are computed. The patch sizes k are chosen such that $k \sim \log(H^{-1})$, that is $k = 1, 2, 2, 3$, and 4 for the five simulations. When computing $U_{H,n}$ the stiffness matrix is assembled on the fine scale h and then interpolated to the coarser scale. This way we avoid quadrature errors. The convergence results for A_1 and A_2 are presented in Figure 2, where the error at the final time t_N is plotted against the degrees of freedom $|\mathcal{N}|$. Comparing the plots we can see the predicted quadratic convergence for the localized GFEM. However, as expected, the P1-FEM shows poor convergence on the coarse grids when the diffusion coefficient has multiscale features. We clearly see the pre-asymptotic effects when H does not resolve the fine structure of B.

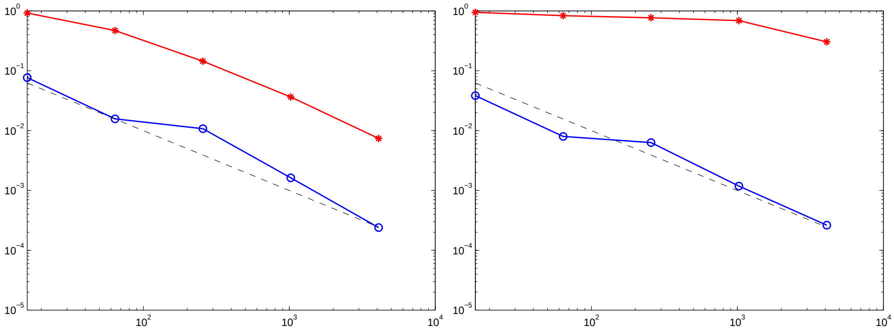


(A) Constant coefficient A_1.

(B) Multiscale coefficient A_2.

Figure 2. Relative L_2 errors $\|U^{\text{ms}}_{k,N} - U_{h,N}\|/\|U_{h,N}\|$ (blue $\circ$) and $\|U_{H,N} - U_{h,N}\|/\|U_{h,N}\|$ (red $*$) for the linear parabolic problem plotted against the number of degrees of freedom $|\mathcal{N}| \approx H^{-2}$. The dashed line is H^2.

6.2. Semilinear parabolic problem. For the semilinear parabolic problem we study the Allen-Cahn equation, which has right hand side $f(u) = -(u^3 - u)$ that fulfills the necessary assumptions. We define the initial data to be $u_0(x,y) = x(1-x)y(1-y)$, which is zero on $\partial\Omega$. The matrix B constructed as in the linear case but with values varying between 10^{-3} and 1. Note that the solution to the Allen-Cahn equation converges to zero rapidly if the diffusion is too high, thus the smaller contrast $\max(B)/\min(B) \approx 10^3$ in this case, see Figure 1 (right). However, B is still rapidly varying. As in the linear case we now compute the localized GFEM approximations $U^{\text{ms}}_{k,n}$ and the corresponding P1-FEM, $U_{H,n}$. The patch sizes are chosen to $k = 1, 2, 2, 3$, and 4, for the five simulations. The convergence results for A_1 and A_2 are presented in Figure 3. We can draw the same conclusions as in the linear case. The localized GFEM shows predicted quadratic convergence in both cases, but P1-FEM shows poor convergence on the coarse grids when the diffusion coefficient has multiscale features.



(A) Constant coefficient A_1.

(B) Multiscale coefficient A_2.

FIGURE 3. Relative L_2 errors $\|U_{k,N}^{\mathrm{ms}} - U_{h,N}\|/\|U_{h,N}\|$ (blue $\circ$) and $\|U_{H,N} - U_{h,N}\|/\|U_{h,N}\|$ (red $*$) for the semilinear parabolic problem plotted against the number of degrees of freedom $|\mathcal{N}| \approx H^{-2}$. The dashed line is H^2.

References

[1] A. Abdulle and P. Henning: *Localized orthogonal decomposition method for the wave equation with a continuum of scales*, Submitted.

[2] I. Babuška and R. Lipton: *Optimal local approximation spaces for generalized finite element methods with application to multiscale problems*, Multiscale Model. Simul. 9 (2011), no. 1, p. 373–406.

[3] R. E. Bank and H. Yserentant: *On the H^1-stability of the L_2-projection onto finite element spaces*, Numer. Math. 126 (2014), no. 2, p. 361–381.

[4] C. Carstensen and R. Verfürth: *Edge residuals dominate a posteriori error estimates for low order finite element methods*, SIAM J. Numer. Anal. 36 (1999), no. 5, p. 1571–1587.

[5] W. E and B. Engquist: *The heterogeneous multiscale methods*, Commun. Math. Sci. 1 (2003), no. 1, p. 87–132.

[6] L. C. Evans: *Partial Differential Equations*, vol. 19 of Graduate Studies in Mathematics, American Mathematical Society, Providence, RI, Second edition, 2010.

[7] P. Henning and A. Målqvist: *Localized orthogonal decomposition techniques for boundary value problems*, SIAM J. Sci. Comput. 36 (2014), no. 4, p. A1609–A1634.

[8] P. Henning, A. Målqvist and D. Peterseim: *A localized orthogonal decomposition method for semi-linear elliptic problems*, ESAIM Math. Model. Numer. Anal. 48 (2014), no. 5, p. 1331–1349.

[9] T. J. R. Hughes, G. R. Feijóo, L. Mazzei, and J-B. Quincy: *The variational multiscale method—a paradigm for computational mechanics*, Comput. Methods Appl. Mech. Engrg. 166 (1998), no. 1-2, p. 3–24.

[10] O. A. Ladyženskaja, V. A. Solonnikov, and N. N. Ural′ceva: *Linear and Quasilinear Equations of Parabolic Type*, Translated from the Russian by S. Smith., Translations of Mathematical Monographs, Vol. 23 American Mathematical Society, Providence, R.I., 1968.

[11] M. G. Larson and A. Målqvist: *Adaptive variational multiscale methods based on a posteriori error estimation: energy norm estimates for elliptic problems*, Comput. Methods Appl. Mech. Engrg. 196 (2007), no. 21-24, p. 2313–2324.

[12] S. Larsson: *Nonsmooth data error estimates with applications to the study of the long-time behavior of finite element solutions of semilinear parabolic problems*, Department of Mathematics, Chalmers University of

Technology, preprint 1992-3, 1992.

[13] S. Larsson: *Semilinear parabolic partial differential equations: theory, approximation, and application*, In *New trends in the mathematical and computer sciences*, vol. 3 of *Publ. ICMCS*, p. 153–194, Int. Cent. Math. Comp. Sci. (ICMCS), Lagos, 2006.

[14] A. Målqvist and D. Peterseim: *Computation of eigenvalues by numerical upscaling*, Numer. Math. 130 (2015), no. 2, p. 337–361.

[15] A. Målqvist and D. Peterseim: *Localization of elliptic multiscale problems*, Math. Comp. 83 (2014), no. 290, p. 2583–2603.

[16] H. Owhadi and L. Zhang: *Homogenization of parabolic equations with a continuum of space and time scales*, SIAM J. Numer. Anal. 46 (2007/08), no. 1, p. 1–36.

[17] H. Owhadi, L. Zhang, and L. Berlyand: *Polyharmonic homogenization, rough polyharmonic splines and sparse super-localization*, ESAIM Math. Model. Numer. Anal. 48 (2014), no. 2, p. 517–552.

[18] V. Thomée: *Galerkin Finite Element Methods for Parabolic Problems*, Springer Series in Computational Mathematics, Springer-Verlag, Berlin, Second edition, 2009.

Paper II

Patrick Henning and Anna Persson, *A multiscale method for linear elasticity reducing Poisson locking*, Preprint.

A multiscale method for linear elasticity reducing Poisson locking

Patrick Henning[1] and Anna Persson[2]

Abstract. We propose a generalized finite element method for linear elasticity equations with highly varying and oscillating coefficients. The method is formulated in the framework of localized orthogonal decomposition techniques introduced by Målqvist and Peterseim [23]. Assuming only L_∞-coefficients we prove linear convergence in the H^1-norm, also for materials with large Lamé parameter λ. The theoretical a priori error estimate is confirmed by numerical examples.

1. Introduction

In this paper we study numerical solutions to linear elasticity equations with highly varying coefficients. Such equations typically occur when modeling the deformation of a heterogeneous material, for instance a composite material. Problems with this type of coefficients are commonly referred to as *multiscale* problems.

The convergence of classical finite element methods based on continuous piecewise polynomials depends on (at least) the spatial H^2-norm of the solution u. However, for problems with multiscale features this norm may be very large. Indeed, if the coefficient varies at a scale of size ϵ, then $\|u\|_{H^2} \sim \epsilon^{-1}$. Thus, to achieve convergence the mesh size must be small ($h < \epsilon$). In many applications this condition leads to issues with computational cost and available memory. To overcome this difficulty several methods have been proposed, where we refer to [1, 9, 24, 29] for multiscale methods particularly addressing elasticity problems.

Generalized finite element methods (GFEM, cf. [4]) belong to the class of Galerkin methods. Instead of constructing the finite dimensional solution space from standard shape functions, a generalized finite element approach is based on constructing a set of locally supported basis functions (not necessarily piecewise polynomials) that incorporate additional information about the structure of the original problem. This strategy can enhance the local approximation properties significantly. In this paper we propose a GFEM based on the ideas in [23], often referred to as *localized orthogonal decomposition* (LOD). The methodology of the LOD arose from the framework of the Variational Multiscale Method (VMM) originally proposed by Hughes et al. [17, 18] as a tool for stabilizing finite element methods that perform bad due to an under-resolution of relevant microscopic data. The stabilization was achieved by using a Petrov-Galerkin formulation of the problem with a standard finite element space as trial space and a generalized finite element space for the test-functions. The concept was

[1] Department of Mathematics, KTH Royal Institute of Technology, SE-100 44 Stockholm, Sweden.

[2] Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, SE-412 96 Göteborg, Sweden.

reinterpreted and specialized in [19, 20] to elliptic homogenization problems. A short time later, the first rigorous analysis was provided in [23] by introducing a H^1-stable localized orthogonal decomposition for constructing the test function space. In subsequent works, refined construction strategies were proposed [16, 13].

The LOD framework relies on a decomposition of a high-dimensional solution space into a coarse space (spanned by a set of standard nodal basis functions) and a fine scale detail space that is expressed through the kernel of a projection operator. The generalized finite element basis functions are constructed by adding a *correction* from the detail space to each coarse nodal basis function. The corrections are problem dependent and constructed by solving a partial differential equation in the fine scale part of the space. In [23] elliptic equations are considered and it is proven that the corrections decay exponentially for these problems. This motivates a truncation to patches of coarse elements, which allow for efficient computations. The resulting method is proved to be convergent of optimal order. This convergence result does not depend on any assumptions regarding periodicity or scale separation of the coefficients. Since its development, the method has been applied to several other types of equations, see, for instance, semilinear elliptic equations [14], boundary value problems [13], eigenvalue problems [22, 15], linear and semilinear parabolic equations [21], the Helmholtz problem [27, 11] and the linear wave equation [2]. A review is given in [28].

In this work we consider linear elasticity equations with mixed inhomogeneous Dirichlet and Neumann boundary conditions. We construct corresponding correctors for standard nodal basis functions and prove that they decay exponentially. Moreover, we prove that the resulting generalized finite element method converges with optimal order in the spatial H^1-norm. The results are confirmed by a numerical example.

Furthermore, the generalized finite element method proposed in this paper reduces the locking effect that is observed for classical finite elements based on continuous piecewise affine polynomials for nearly incompressible materials. The error bound derived for the ideal method (without localization) is uniform in the Lamé parameter λ, i.e., completely locking-free. The error estimate for the final localized method depends on λ, however not in the usual manner, but only weakly through a term that converges with an exponential rate to zero. In practice, this eliminates the locking-effect.

The paper is organized as follows. In Section 2 we formulate the problem, in Section 3 we define the generalized finite element method and in Section 4 we perform the localization of the basis functions. Finally, in Section 5 we provide some numerical examples.

2. Problem formulation

Let $d = 2, 3$, denote the spatial dimension and let $\mathbb{S} := \mathbb{R}^{d\times d}_{\text{sym}}$ denote the space of $d \times d$ symmetric matrices over $\mathbb{R}$. On $\mathbb{S}$, we use the double-dot product notation

$$A : B = \sum_{i,j=1}^{d} A_{ij} B_{ij}, \quad A, B \in \mathbb{S}.$$

The computational domain $\Omega \subseteq \mathbb{R}^d$ is assumed to be a bounded polygonal (or polyhedral) Lipschitz domain describing the reference configuration of an elastic medium. We use $(\cdot,\cdot)_{L_2(\Omega)}$ to denote the inner product on $L_2(\Omega, \mathbb{R}^d)$

$$(v, w)_{L_2(\Omega)} := \int_\Omega v(x) \cdot w(x) \, \mathrm{d}x, \quad v, w \in L_2(\Omega, \mathbb{R}^d),$$

and $\|\cdot\|_{L_2(\Omega)}$ for the corresponding norm. Furthermore, we let $H^1(\Omega, \mathbb{R}^d)$ denote the classical Sobolev space with norm $\|v\|^2_{H^1(\Omega)} := \|v\|^2_{L_2(\Omega)} + \|\nabla v\|^2_{L_2(\Omega)}$, where $\nabla v \in L_2(\Omega, \mathbb{R}^{d \times d})$, and

$$\|\nabla v\|^2_{L_2(\Omega)} := \sum_{i,j=1}^{d} \int_\Omega (\partial_i v_j(x))^2 \, \mathrm{d}x, \quad v \in H^1(\Omega, \mathbb{R}^d).$$

Let $u : \Omega \to \mathbb{R}^d$ denote the displacement field of the elastic medium. Under the assumption of small displacement gradients, the (linearized) strain tensor $\varepsilon(u)$ is given by

$$\varepsilon_{kl}(u) := \frac{1}{2}(\partial_k u_l + \partial_l u_k), \quad 1 \leq k, l \leq d.$$

Furthermore, Hooke's (generalized) law states that the stress tensor σ is given by the relation

$$\sigma_{ij} = \sum_{k,l=1}^{d} A_{ijkl}(x) \varepsilon_{kl}(u), \quad 1 \leq i, j \leq d,$$

where A is a fourth order tensor describing the elastic medium. In this paper we assume that the material is strongly heterogeneous and thus A has multiscale properties. The tensor A is assumed to be symmetric in the sense that $A_{ijkl} = A_{jikl} = A_{ijlk} = A_{klij}$ almost everywhere.

Cauchy's equilibrium equation now states that

$$-\nabla \cdot \sigma = f,$$

where $f : \Omega \to \mathbb{R}^d$ denotes the body forces. To formulate the problem of interest we let Γ_D and Γ_N denote two disjoint Hausdorff measurable segments of the boundary, such that $\Gamma_D \cup \Gamma_N = \partial\Omega$, where Dirichlet and Neumann conditions are imposed respectively. The linear elasticity problem consists of finding the displacement u and the stress tensor σ such that

$$-\nabla \cdot \sigma = f, \qquad \text{in } \Omega, \tag{2.1}$$

$$\sigma_{ij} = \sum_{k,l=1}^{d} A_{ijkl}\,\varepsilon_{kl}(u), \quad \text{in } \Omega, \tag{2.2}$$

$$u = g, \qquad \text{on } \Gamma_D, \tag{2.3}$$

$$\sigma \cdot n = b, \qquad \text{on } \Gamma_N, \tag{2.4}$$

where we assume that $\text{meas}(\Gamma_D) > 0$. Here $g, b : \Omega \to \mathbb{R}^d$ denotes the Dirichlet and Neumann data respectively.

To pose a variational form of problem (2.1)-(2.4) we need to define appropriate test and trial spaces. Letting $\gamma : H^1(\Omega) \to L_2(\Gamma_D)$ denote the trace operator onto Γ_D, we define the test space

$$V := \{v \in (H^1(\Omega))^d : \gamma v = 0\}.$$

Multiplying the equation (2.1) with a test function from V and using Green's formula together with the boundary conditions (2.4) we get that

$$(\sigma : \nabla v)_{L_2(\Omega)} = (f, v)_{L_2(\Omega)} + (b, v)_{L_2(\Gamma_N)}.$$

Due to the symmetry of A we have the identity $(\sigma : \nabla v) = (\sigma : \varepsilon(v))$, and by defining the bilinear form

$$\mathcal{B}(u, v) := (\sigma : \varepsilon(v))_{L_2(\Omega)} = (A(x)\varepsilon(u) : \varepsilon(v))_{L_2(\Omega)},$$

we arrive at the following weak formulation of (2.1)-(2.4). Find $u \in H^1(\Omega, \mathbb{R}^d)$, such that $\gamma u = g$, and

$$\mathcal{B}(u, v) = (f, v)_{L_2(\Omega)} + (b, v)_{L_2(\Gamma_N)}, \quad \forall v \in V. \tag{2.5}$$

Remark 2.1. In the case of an isotropic medium the elasticity coefficient satisfies $A_{ijkl} = \mu(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}) + \lambda\delta_{ij}\delta_{kl}$, where δ_{ij} is the Kronecker delta, and μ and λ are the so called Lamé coefficients. The stress tensor can in this case be simplified to

$$\sigma = 2\mu\varepsilon(u) + \lambda(\nabla \cdot u)I,$$

where I is the identity matrix.

Assumptions. We make the following assumptions on the data

(A1) $A_{ijkl} \in L_\infty(\Omega, \mathbb{R})$, $1 \leq i, j, k, l \leq d$, and there exist positive constants $\alpha, \beta \in \mathbb{R}$ such that

$$\alpha B : B \leq A(\cdot)B : B \leq \beta B : B, \quad \forall B \in \mathbb{S}, \quad \text{a.e. in } \Omega.$$

(A2) $f \in L_2(\Omega, \mathbb{R}^d)$, $b \in L_2(\Gamma_N, \mathbb{R}^d)$, and $g \in H^{1/2}(\Gamma_D, \mathbb{R}^d)$.

Recall Korn's inequality for a domain with mixed boundary conditions, see, for instance, [7, 25].

Lemma 2.2 (Korn's inequality). *Let $\Omega \subset \mathbb{R}^d$ denote a bounded and connected Lipschitz-domain, and let Γ_D denote the part of the boundary where Dirichlet boundary conditions are defined. If* $\operatorname{meas}(\Gamma_D) > 0$*, then*

$$\|\nabla v\|_{L_2(\Omega)} \leq C_{\text{ko}} \|\varepsilon(v)\|_{L_2(\Omega)}, \quad \forall v \in V, \tag{2.6}$$

Here C_{ko} is a constant depending only on Ω.

In the case $\Gamma_D = \partial\Omega$ we have $C_{\text{ko}} = \sqrt{2}$, independently of the size of Ω. Using (2.6) we derive the following bounds,

$$\alpha C_{\text{ko}}^{-2} \|\nabla v\|^2_{L_2(\Omega)} \leq \mathcal{B}(v,v) \leq \beta \|\nabla v\|^2_{L_2(\Omega)}, \quad \forall v \in V, \tag{2.7}$$

where we have used the bound $\|\varepsilon(v)\|_{L_2(\Omega)} \leq \|\nabla v\|_{L_2(\Omega)}$. It follows that the bilinear form $\mathcal{B}(\cdot,\cdot)$ is an inner product on V and existence and uniqueness of a solution to the problem (2.5) follows from the Lax-Milgram lemma. We denote the norm induced by the inner product $\mathcal{B}(\cdot,\cdot)$ by $\|v\|^2_{\mathcal{B}(\Omega)} := \mathcal{B}(v,v)$ for $v \in V$.

Remark 2.3. In the case of an isotropic material (see Remark 2.1) we have the bounds

$$\begin{aligned} C_{\text{ko}}^{-2} 2\mu_1 \|\nabla v\|^2_{L_2(\Omega)} &\leq \|\sqrt{2\mu}\varepsilon(v)\|^2_{L_2(\Omega)} \leq \|\sqrt{2\mu}\varepsilon(v)\|^2_{L_2(\Omega)} + \|\sqrt{\lambda}\nabla \cdot v\|^2_{L_2(\Omega)} \\ &= \mathcal{B}(v,v) \leq C(2\mu_2 + \lambda_2)\|\nabla v\|^2_{L_2(\Omega)}, \end{aligned}$$

where $\mu_1 > 0$ is the lower bound of μ and $\mu_2, \lambda_2 \leq \infty$ are the upper bounds of μ and λ respectively. We emphasize that this means that only β in (2.7) depends on λ.

3. Numerical Approximation

3.1. Classical finite element. First, we define the classical finite element space of continuous and piecewise affine elements. Let $\mathcal{T}_h$ be a regular triangulation of Ω into closed triangles/tetrahedra with mesh size $h_T := \operatorname{diam}(T)$, for $T \in \mathcal{T}_h$, and denote the largest diameter in the triangulation by $h := \max_{T \in \mathcal{T}_h} h_T$. We assume that the family of triangulations $\{\mathcal{T}_h\}_{h>0}$ is shape regular. Now define the spaces

$$\begin{aligned} &S_h = \{v \in (C(\bar{\Omega}))^d : v_j|_T \text{ is a polynomial of degree } \leq 1, \forall T \in \mathcal{T}_h, 1 \leq j \leq d\}, \\ &V_h = S_h \cap V. \end{aligned}$$

Furthermore, we let $\mathcal{N}_h$ denote the nodes generated by $\mathcal{T}_h$ and $\mathring{\mathcal{N}}_h = \mathcal{N}_h \setminus \Gamma_D$ the free nodes in V_h. Now, let $g_h \in S_h$ be an approximation of an extension of g, such that $g_h(z) = 0$, $\forall z \in \mathring{\mathcal{N}}_h$ and γg_h is some appropriate approximation of g. The classical finite element method now reads; find $u_h = u_{h,0} + g_h$, such that $u_{h,0} \in V_h$ and

$$\mathcal{B}(u_{h,0}, v) = (f,v)_{L_2(\Omega)} + (b,v)_{L_2(\Gamma_N)} - \mathcal{B}(g_h, v), \quad \forall v \in V_h. \tag{3.1}$$

Note that $\gamma u_h = \gamma g_h$, where γg_h is an approximation of g.

Theorem 3.1. *Let u be the solution to* (2.5) *and u_h the solution to* (3.1). *If the solution u is sufficiently regular we have*

$$\|u - u_h\|_{H^1(\Omega)} \leq C_A h \|D^2 u\|_{L_2(\Omega)},$$

where C_A depends on the size of A and $\|D^2u\|_{L_2(\Omega)}$ depends on the variations in A via a regularity estimate $\|D^2u\|_{L_2(\Omega)} \leq C(u,\Omega)\|A\|_{W^{1,\infty}(\Omega)}$. In particular, we have $\|D^2u\|_{L_2(\Omega)} \to \infty$ the faster A oscillates.

Since the a priori bound in Theorem 3.1 depends, through the H^2-norm of u, on the variations (derivatives) in the data, the mesh width h must be sufficiently small for u_h to be a good approximation of u. In the context of multiscale problems, this results in a significant computational complexity. In the following we assume that h is small enough and we shall refer to u_h as a reference solution. However, we emphasize that our method never requires to compute this expensive reference solution and that it is purely used for comparisons.

3.1.1. *Poisson locking.* This subsection describes the phenomenon known as locking, sometimes referred to as Poisson locking to distinguish it from other types of locking. To simplify the discussion here we assume that we have an isotropic material with μ and λ *constant* parameters and $g_D = 0$ on $\Gamma_D = \partial\Omega$. In this case we can exploit Galerkin orthogonality and the norm-equivalence in Remark 2.1 to see that the error bound in Theorem 3.1 becomes the estimate

$$\|u - u_h\|_{H^1(\Omega)} \leq Ch\frac{\sqrt{2\mu + \lambda}}{\sqrt{2\mu}}\|D^2u\|_{L_2(\Omega)}, \tag{3.2}$$

where C is independent of μ and λ. Moreover, $\|D^2u\|_{L_2(\Omega)}$ is independent of μ and λ which follows from the stability estimate (see [8]),

$$\|u\|_{H^2(\Omega)} + \lambda\|\nabla \cdot u\|_{H^1(\Omega)} \leq C_\Omega\|f\|_{L_2(\Omega)}, \tag{3.3}$$

where C_Ω is independent of μ and λ. We emphasize that the estimate (3.3) does not hold if μ and λ vary in space. Since both C and $\|D^2u\|_{L_2(\Omega)}$ in (3.2) are independent of λ, we conclude that the error bound blows up as $\lambda \to \infty$. This is counter-intuitive to the observation that the error with respect to the H^1-best-approximation in V_h is not affected by λ.

In fact, there is a simple reason for this phenomenon. For $\lambda \to \infty$ we have that the displacement must fulfill the extra condition $\nabla \cdot u = 0$. However, $v_h = 0$ is the only function in V_h that fulfills $\nabla \cdot v_h = 0$. This forces the Galerkin-approximation u_h to convergence to the bad approximation $u_h = 0$ in order to remain stable. This issue can be avoided by using discrete solution spaces in which divergence-free functions can be well-approximated, cf. the robust methods in [7, 8, 5, 3], where it is in fact possible to derive estimates of the type $\|u - u_h\|_{H^1(\Omega)} \leq Ch\|D^2u\|_{L_2(\Omega)}$ independent of λ.

From the discussion above we conclude that if λ is large compared to μ the mesh size must be sufficiently small, i.e. $h \lesssim 1/\sqrt{\lambda}$, to achieve convergence for conventional Lagrange $P1$ finite elements. A natural question is what the typical ranges of values for μ and λ are and how they are related. The Lamé parameters are determined by Young's modulus E and Poisson's ratio ν according to $\mu =$

$\frac{E}{2(1+\nu)}$ and $\lambda = \frac{E\nu}{(1+\nu)(1-2\nu)}$. Consequently, we obtain $\frac{\sqrt{2\mu+\lambda}}{\sqrt{2\mu}} = \sqrt{\frac{1}{1-2\nu}}$ and hence (3.2) reduces to

$$\|u - u_h\|_{H^1(\Omega)} \leq C_\Omega \frac{h}{\sqrt{1-2\nu}} \|f\|_{L_2(\Omega)}, \tag{3.4}$$

where we see that the problem only arises if the Poisson's ratio is close to $\nu = 0.5$, which describes a perfectly incompressible material. In most engineering applications the value of Poisson's ratio lies between 0.2 and 0.35 (e.g. $\nu = 0.27 - 0.30$ for steel, $\nu = 0.2 - 0.3$ for rocks such as granite or sandstone and $\nu = 0.17 - 0.27$ for glass; cf. [12]). Poisson's ratios larger than 0.45 are rare. Examples for such tough cases are clay ($\nu \leq 0.45$), gold ($\nu = 0.45$) and lead ($\nu = 0.46$). Natural rubber with $\nu = 0.4999$ can be considered as the most extreme case (cf. [26]). These values give us a clear image about the order of magnitude required for h in practical scenarios. If the extension of Ω is of order 1, tough cases ($\nu \approx 0.45$) require $h \lesssim \frac{1}{3}$ and extreme cases ($\nu \approx 0.4999$) require $h \lesssim \frac{1}{70}$. These values help us to understand the phenomenon of locking better. The constraints that are imposed by Poisson locking are not severe (in the sense that it does typically not make the problem prohibitively expensive), but they are highly impractical and not desirable in the sense that they make the problem significantly more expansive than it should be. For instance for $\nu = 0.45$ the mesh needs to be three times finer than for a locking-free method, which makes an enormous difference in CPU demands due to the curse of dimension.

3.1.2. *Poisson locking for multiscale problems.* This paper is devoted to multiscale problems and the locking effect has to be seen from a different perspective in this case. Multiscale elasticity problems as they typically arise in engineering or in geosciences involve material parameters (in general form represented by the tensor $A(x)$) that vary on an extremely fine scale ϵ (relative to the extension of the computational domain) with $\epsilon \ll \lambda^{-1/2}$. These variations need to be resolved by an underlying fine mesh which imposes the condition $h < \epsilon \ll \lambda^{-1/2}$ even for locking-free methods. In other words, the natural constraints imposed by the variations of the coefficient are much more severe than the constraints imposed by the locking effect. Since we assume that the reference solution u_h given by (3.1) is a good approximation to our original multiscale problem (i.e. $h < \epsilon$), then the solution will not suffer from the locking effect either. For that reason we consider u_h as being locking-free. Our multiscale method is constructed to approximate u_h on significantly coarser scales of order H, and we call this method a *locking-free multiscale method* if the convergence rates in H are independent of λ and the variations of A.

Locking and multiscale are two different characteristics that typically need to be treated with different approaches, as a multiscale method is not necessarily locking-free. In the following we show that the framework of the LOD can be used for stabilizing $P1$ Lagrange finite elements in such a way that both effects are reduced simultaneously. In particular we show that it is not necessary to use higher order Lagrange elements, discontinuous Galerkin approaches, mixed finite elements or Crouzeix-Raviart finite elements as they are commonly required for eliminating Poisson locking.

In this paper the error estimate for the ideal method (without localization) in Lemma 3.2 is independent of λ and thus locking-free. The localization depends on the contrast β/α, see Theorem 4.1. However, this ratio enters only through a term that converges with exponential order to zero. Consequently, the locking effect decays exponentially in the localized method. This is also tested numerically in Section 5.

3.2. Generalized finite element. In this subsection we introduce a generalized finite element method. Let V_H denote the same classical finite element space as V_h, but with a coarser mesh size $H > h$. Let $\mathcal{T}_H$ be the triangulation associated with the space V_H and assume that $\mathcal{T}_h$ is a refinement of $\mathcal{T}_H$ such that $V_H \subseteq V_h$. In addition to shape regular, we assume the family $\{\mathcal{T}_H\}_{H>h}$ to be quasi-uniform.

We define $\mathcal{N}_H$ and $\mathring{\mathcal{N}}_H$ analogously to $\mathcal{N}_h$ and $\mathring{\mathcal{N}}_h$. Note that the mesh width H is too coarse for the classical finite element solution (3.1) in V_H to be a good approximation. The aim is now to define a new (multiscale) space with the same dimension as V_H, but with better approximation properties.

To define such a multiscale space we need to introduce some notation. First, let $I_H : V_h \to V_H$ denote an interpolation operator with the property that $I_H \circ I_H = I_H$ and

$$H_T^{-1}\|v - I_H v\|_{L_2(T)} + \|\nabla I_H v\|_{L_2(T)} \leq C_I \|\nabla v\|_{L_2(\omega_T)}, \quad \forall T \in \mathcal{T}_H,\ v \in V_h, \tag{3.5}$$

where

$$\omega_T := \cup\{\hat{T} \in \mathcal{T}_H : \hat{T} \cap T \neq \emptyset\}.$$

For a shape regular mesh, the estimates in (3.5) can be summed to a global estimate

$$H^{-1}\|v - I_H v\|_{L_2(\Omega)} + \|\nabla I_H v\|_{L_2(\Omega)} \leq C_\rho \|\nabla v\|_{L_2(\Omega)}, \tag{3.6}$$

where C_ρ depends on C_I and the shape regularity parameter, $\rho > 0$;

$$\rho := \max_{T\in\mathcal{T}_H} \rho_T, \text{ with } \rho_T := \frac{\operatorname{diam} B_T}{\operatorname{diam} T}, \text{ for } T \in \mathcal{T}_H.$$

Here B_T is the largest ball contained in T. For instance, we could choose $I_H^i = E_H^i \circ \Pi_H^i$, $1 \leq i \leq d$, where Π_H^i is the L_2-projection onto $P_1(\mathcal{T}_H)$, the space of functions that are affine on each triangle $T \in \mathcal{T}_H$ and $E_H^i : P_1(\mathcal{T}_H) \to V_H$ the averaging operator defined by

$$(E_H^i(v))(z) = \frac{1}{\operatorname{card}\{T \in \mathcal{T}_H : z \in T\}} \sum_{T\in\mathcal{T}_H : z\in T} v|_T(z), \tag{3.7}$$

where $z \in \mathring{\mathcal{N}}_H$, see [28] for further details and other possible choices of I_H.

Let V_{f} denote the kernel to the operator I_H

$$V_{\mathrm{f}} := \ker I_H = \{v \in V_h : I_H v = 0\}.$$

The space V_h can now be split into the two spaces $V_h = V_H \oplus V_{\mathrm{f}}$, meaning that $v_h \in V_h$ can be decomposed into $v_h = v_H + v_{\mathrm{f}}$, such that $v_H \in V_H$ and $v_{\mathrm{f}} \in V_{\mathrm{f}}$. The kernel V_{f} is a detail space in the sense that it captures all features that are not captured by the (coarse) space V_H.

Let $R_{\mathrm{f}} : V_h \to V_{\mathrm{f}}$ be the Ritz projection onto V_{f} using the inner product $\mathcal{B}(\cdot,\cdot)$ such that

$$\mathcal{B}(R_{\mathrm{f}} v, w) = \mathcal{B}(v, w), \quad \forall w \in V_{\mathrm{f}}, \quad v \in V_h. \tag{3.8}$$

Since $v_h = v_H + v_{\mathrm{f}}$ with $v_H \in V_H$ and $v_{\mathrm{f}} \in V_{\mathrm{f}}$ we have

$$v_h - R_{\mathrm{f}} v_h = v_H - R_{\mathrm{f}} v_H, \quad \forall v_h \in V_h,$$

and we define the multiscale space

$$V_{\mathrm{ms}} = \{v_H - R_{\mathrm{f}} v_H : v_H \in V_H\}. \tag{3.9}$$

Note that this space has the same dimension as V_H, but contains fine scale features. Indeed, with λ_z denoting the hat basis function in V_H corresponding to node z, the set

$$\{\lambda_z - R_{\mathrm{f}} \lambda_z : z \in \mathring{\mathcal{N}}_H\},$$

is a basis for V_{ms}. Moreover, we note that V_{ms} is the orthogonal complement to V_{f} with respect to the inner product $\mathcal{B}(\cdot,\cdot)$. Thus the split $V_h = V_{\mathrm{ms}} \oplus V_{\mathrm{f}}$ and the following orthogonality holds for $v_{\mathrm{ms}} \in V_{\mathrm{ms}}$ and $v_{\mathrm{f}} \in V_{\mathrm{f}}$

$$\mathcal{B}(v_{\mathrm{ms}}, v_{\mathrm{f}}) = \mathcal{B}(v_{\mathrm{f}}, v_{\mathrm{ms}}) = 0. \tag{3.10}$$

To define a generalized finite element method we aim to replace the space V_h with V_{ms} in (3.1). Due to the inhomogeneous boundary conditions we also need two extra corrections similar to the ones used in [13]. For the Dirichlet condition we subtract $R_{\mathrm{f}} g_h$ from the solution. For the Neumann condition we define a correction $\tilde{b}_{\mathrm{f}} \in V_{\mathrm{f}}$ such that

$$\mathcal{B}(\tilde{b}_{\mathrm{f}}, w) = (b, w)_{L_2(\Gamma_N)}, \quad \forall w \in V_{\mathrm{f}}. \tag{3.11}$$

We are now ready to define the generalized finite element method; find

$$u_{\mathrm{ms}} = u_{0,\mathrm{ms}} + \tilde{b}_{\mathrm{f}} + g_h - R_{\mathrm{f}} g_h,$$

such that $u_{0,\mathrm{ms}} \in V_{\mathrm{ms}}$ and

$$\mathcal{B}(u_{0,\mathrm{ms}}, v) = (f, v)_{L_2(\Omega)} + (b, v)_{L_2(\Gamma_N)} - \mathcal{B}(\tilde{b}_{\mathrm{f}} + g_h - R_{\mathrm{f}} g_h, v), \quad \forall v \in V_{\mathrm{ms}}. \tag{3.12}$$

Note that both $\tilde{b}_{\mathrm{f}} = R_{\mathrm{f}} g_h = 0$ on Γ_D, so $\gamma u_{\mathrm{ms}} = \gamma g_h$, and

$$\mathcal{B}(u_{\mathrm{ms}}, v) = (f, v)_{L_2(\Omega)} + (b, v)_{L_2(\Gamma_N)}, \quad \forall v \in V_{\mathrm{ms}},$$

as desired.

Lemma 3.2. *Let u_h be the solution to* (3.1) *and u_{ms} the solution to* (3.12). *Then*

$$\|u_h - u_{\mathrm{ms}}\|_{H^1(\Omega)} \leq CH\alpha^{-1}\|f\|_{L_2(\Omega)}, \tag{3.13}$$

where C depends on C_{ko} and C_ρ.

Proof. Define $e := u_h - u_{\mathrm{ms}}$. Since $V_{\mathrm{ms}} \subseteq V_h$, we have the Galerkin orthogonality

$$\mathcal{B}(e, v) = 0, \quad \forall v \in V_{\mathrm{ms}}.$$

Recall that we can write $e = (I - R_{\mathrm{f}})e + R_{\mathrm{f}}e$ where $(I - R_{\mathrm{f}})e \in V_{\mathrm{ms}}$ and $R_{\mathrm{f}}e \in V_{\mathrm{f}}$. Using this we get

$$\begin{aligned}\alpha C_{\mathrm{ko}}^{-2}\|\nabla e\|^2_{L_2(\Omega)} &\leq \mathcal{B}(e,e) = \mathcal{B}(e, R_{\mathrm{f}}e) = \mathcal{B}(u_h - u_{\mathrm{ms}}, R_{\mathrm{f}}e)\\ &= (f, R_{\mathrm{f}}e)_{L_2(\Omega)} + (b, R_{\mathrm{f}}e)_{L_2(\Gamma_N)}\\ &\qquad - \mathcal{B}(u_{0,\mathrm{ms}} + \tilde{b}_{\mathrm{f}} + g_h - R_{\mathrm{f}}g_h, R_{\mathrm{f}}e)\\ &= (f, R_{\mathrm{f}}e)_{L_2(\Omega)},\end{aligned}$$

where have used the orthogonality (3.10) and the definitions (3.11) and (3.8) in the last equality. Now, since $R_{\mathrm{f}}e \in V_{\mathrm{f}}$ we have that $I_H R_{\mathrm{f}}e = 0$ and using (3.6) we get

$$\begin{aligned}\alpha C_{\mathrm{ko}}^{-2}\|\nabla e\|^2_{L_2(\Omega)} &\leq \mathcal{B}(e,e) \leq (f, R_{\mathrm{f}}e - I_H R_{\mathrm{f}}e)_{L_2(\Omega)}\\ &\leq \|f\|_{L_2(\Omega)}\|R_{\mathrm{f}}e - I_H R_{\mathrm{f}}e\|_{L_2(\Omega)} \leq C_\rho H\|f\|_{L_2(\Omega)}\|\nabla e\|_{L_2(\Omega)},\end{aligned} \tag{3.14}$$

and (3.13) follows. □

4. Localization

The problem of finding $R_{\mathrm{f}}\lambda_z$ in (3.9) is posed in the entire fine scale space V_{f} and thus computationally expensive. Moreover, the resulting basis functions may have global support. However, as we show in this section, the basis functions have exponential decay away from node z, which motivates a truncation of the basis functions. This truncation significantly reduces the computational cost and the resulting functions have local support.

We consider a localization strategy similar to the one proposed in [13]. We restrict the fine scale space V_{f} to patches $\omega_k(T)$ of coarse elements of the following type; for $T \in \mathcal{T}_H$

$$\begin{aligned}&\omega_0(T) := \operatorname{int} T,\\ &\omega_k(T) := \operatorname{int}\big(\cup\{\hat{T} \in \mathcal{T}_H : \hat{T} \cap \overline{\omega_{k-1}(T)} \neq \emptyset\}\big), \quad k = 1, 2, ...\end{aligned}$$

Define $V_{\mathrm{f}}(\omega_k(T)) := \{v \in V_{\mathrm{f}} : v = 0 \text{ on } (\overline{\Omega} \setminus \Gamma_N) \setminus \omega_k(K)\}$ to be the restriction of V_{f} to the patch $\omega_k(T)$. Note that the functions in $V_{\mathrm{f}}(\omega_k(T))$ are zero on the boundary $\partial\omega_k(T) \setminus \Gamma_N$.

We proceed by noting that the Ritz projection R_{f} in (3.8) can be written as the sum

$$R_{\mathrm{f}} = \sum_{T \in \mathcal{T}_H} R_{\mathrm{f}}^T,$$

where $R_{\mathrm{f}}^T : V_h \to V_{\mathrm{f}}$ and fulfills

$$\mathcal{B}(R_{\mathrm{f}}^T v, w) = \mathcal{B}(v, w)_T, \quad \forall w \in V_{\mathrm{f}}, \quad v \in V_h, \quad T \in \mathcal{T}_H, \tag{4.1}$$

where we define

$$\mathcal{B}(v, w)_T := (A\varepsilon(v) : \varepsilon(w))_{L_2(T)}, \quad T \in \mathcal{T}_H.$$

We now aim to localize these computations by replacing V_{f} with $V_{\mathrm{f}}(\omega_k(T))$. Define $R_{\mathrm{f},k}^T : V_h \to V_{\mathrm{f}}(\omega_k(T))$ such that

$$\mathcal{B}(R_{\mathrm{f},k}^T v, w) = \mathcal{B}(v, w)_T, \quad \forall w \in V_{\mathrm{f}}(\omega_k(T)), \quad v \in V_h, \quad T \in \mathcal{T}_H, \tag{4.2}$$

and set $R_{\mathrm{f},k} := \sum_{T \in \mathcal{T}_H} R_{\mathrm{f},k}^T$. We can now define the localized multiscale space

$$V_{\mathrm{ms},k} = \{v_H - R_{\mathrm{f},k} v_H : v_H \in V_H\}. \tag{4.3}$$

Using the same techniques we also define localized versions of the Neumann boundary correctors (3.11). Note that $\tilde{b}_{\mathrm{f}} = \sum_{T \in \mathcal{T}_H \cap \Gamma_N} \tilde{b}_{\mathrm{f}}^T$ where $\tilde{b}_{\mathrm{f}}^T$ is defined by

$$\mathcal{B}(\tilde{b}_{\mathrm{f}}^T, w) = (b, w)_{L_2(\Gamma_N \cap T)}, \quad \forall w \in V_{\mathrm{f}}, \quad T \in \mathcal{T}_H, \quad T \cap \Gamma_N \neq \emptyset,$$

Thus, we define $\tilde{b}_{\mathrm{f},k}^T \in V_{\mathrm{f}}(\omega_k(T))$ such that

$$\mathcal{B}(\tilde{b}_{\mathrm{f},k}^T, w) = (b, w)_{L_2(\Gamma_N \cap T)}, \quad \forall w \in V_{\mathrm{f}}(\omega_k(T)), \quad T \in \mathcal{T}_H, \quad T \cap \Gamma_N \neq \emptyset,$$

and set $\tilde{b}_{\mathrm{f},k} = \sum_{T \in \mathcal{T}_H} \tilde{b}_{\mathrm{f},k}^T$.

We are now ready to define a localized version of (3.12); find

$$u_{\mathrm{ms},k} = u_{0,\mathrm{ms},k} + \tilde{b}_{\mathrm{f},k} + g_h - R_{\mathrm{f},k} g_h,$$

such that $u_{0,\mathrm{ms},k} \in V_{\mathrm{ms},k}$ and

$$\begin{aligned}\mathcal{B}(u_{0,\mathrm{ms},k}, v) \\ = (f, v)_{L_2(\Omega)} + (b, v)_{L_2(\Gamma_N)} - \mathcal{B}(\tilde{b}_{\mathrm{f},k} + g_h - R_{\mathrm{f},k} g_h, v), \quad \forall v \in V_{\mathrm{ms},k}.\end{aligned} \tag{4.4}$$

As for the non-localized problem (3.12), we note that $\tilde{b}_{\mathrm{f},k}$ and $R_{\mathrm{f},k}$ vanish on Γ_D, so $\gamma u_{\mathrm{ms},k} = \gamma g_h$, and

$$\mathcal{B}(u_{\mathrm{ms},k}, v) = (f, v)_{L_2(\Omega)} + (b, v)_{L_2(\Gamma_N)}, \quad \forall v \in V_{\mathrm{ms},k}.$$

The main result in this paper is the following theorem.

Theorem 4.1. *Let u_h be the solution to* (3.1) *and $u_{\mathrm{ms},k}$ the solution to* (3.12)*. Then there exists $\theta \in (0,1)$, depending on the contrast β/α, such that*

(4.5)

$$\begin{aligned}\|u_h - u_{\mathrm{ms},k}\|_{H^1(\Omega)} \leq & CH\alpha^{-1}\|f\|_{L_2(\Omega)} \\ & + Ck^{d/2}\theta^k\sqrt{\frac{\beta^3}{\alpha^5}}(\|f\|_{L_2(\Omega)} + \|b\|_{L_2(\Gamma_N)} + \sqrt{\alpha}\|g_h\|_{\mathcal{B}(\Omega)}),\end{aligned}$$

where C and θ depends on C_{ko}, ρ, and C_I, but not on k, h, H, or the variations of A.

To prove the a priori bound in Theorem 4.1 we first prove three lemmas. In the proofs we use the cut-off functions $\eta_k^T \in V_H$ with nodal values

$$\eta_k^T(x) = 0, \quad \forall x \in \mathcal{N} \cap \overline{\omega_{k-1}(T)}, \tag{4.6a}$$

$$\eta_k^T(x) = 1, \quad \forall x \in \mathcal{N} \cap (\Omega \setminus \omega_k(T)). \tag{4.6b}$$

These functions satisfy the following Lipschitz bound

$$\|\nabla \eta_k^T\|_{L_\infty(\Omega)} \leq CH^{-1}, \quad T \in \mathcal{T}_H, \tag{4.7}$$

where C now depends on the quasi-uniformity. The proof technique relies on the multiplication of a function in the fine scale space V_{f} with a cut-off function. However, this product does not generally belong to the space V_{f}. To fix this, let $\mathcal{I}_h : V \to V_h$ denote the classical linear Lagrange interpolation onto V_h. Using that I_H in (3.7) is a projection we get

$$z := (I - I_H)\mathcal{I}_h(\eta_k^T w) \in V_{\mathrm{f}}(\Omega \setminus \omega_{k-2}(T)), \quad \forall w \in V_{\mathrm{f}},$$

where I denotes the identity mapping. Note that the Lagrange interpolation is needed since $\eta_k^T w \notin V_h$. Furthermore, we have $\operatorname{supp} \mathcal{I}_h(\eta_k^T w) \subseteq \Omega \setminus \omega_{k-1}(T)$ and $\operatorname{supp} I_H\mathcal{I}_h(\eta_k^T R_{\mathrm{f}}^T v) \subseteq \Omega \setminus \omega_{k-2}(T)$ and we conclude $z \in V_{\mathrm{f}}(\Omega \setminus \omega_{k-2}(T))$.

Lemma 4.2. *For $w \in V_{\mathrm{f}}$ and $z := (I - I_H)\mathcal{I}_h\eta_k^T w \in V_{\mathrm{f}}(\Omega \setminus \omega_{k-2}(T))$ it holds that* $\operatorname{supp}(w - z) \subseteq \omega_k(T)$ *and*

$$\|\nabla(w-z)\|_{L_2(\omega_k(T)\setminus\omega_{k-2}(T))} \leq C_{I,\eta}\|\nabla w\|_{L_2(\omega_{k+1}(T)\setminus\omega_{k-3}(T))}, \tag{4.8}$$

$$\|\nabla(w-z)\|_{L_2(\omega_k(T))} \leq C'_{I,\eta}\|\nabla w\|_{L_2(\omega_{k+1}(T))}, \tag{4.9}$$

$$\|\nabla z\|_{L_2(\Omega\setminus\omega_{k-2}(T))} \leq C''_{I,\eta}\|\nabla w\|_{L_2(\Omega\setminus\omega_{k-3}(T))}, \tag{4.10}$$

where $C_{I,\eta}$, $C'_{I,\eta}$, and $C''_{I,\eta}$ depends on C_I, ρ, and the bound in (4.7)*, but not on k, h, H, T, or the variations of A.*

Proof. We have $\eta_k^T = 1$ on $\Omega \setminus \omega_k(T)$ and hence

$$w - z = w - (I - I_H)w = 0, \quad \text{on } \Omega \setminus \omega_k(T),$$

since $I_H w = 0$ and it follows that $\operatorname{supp}(w - z) \subseteq \omega_k(T)$.

Now, note that

$$w - z = (I - I_H)(w - \mathcal{I}_h(\eta_k^T w)).$$

Using the stability of I_H in (3.5) we derive the bound

$$\begin{aligned}\|\nabla(I-I_H)(w-\mathcal{I}_h(\eta_k^T w))\|_{L_2(\omega_k(T)\setminus\omega_{k-2}(T))}&\\ \leq C_I\|\nabla(w-\mathcal{I}_h(\eta_k^T w))\|_{L_2(\omega_{k+1}(T)\setminus\omega_{k-3}(T))}&.\end{aligned}$$

Now, using that the Lagrange interpolation $\mathcal{I}_h$ is H^1-stable for piecewise second order polynomials on shape regular meshes and the bound (4.7) we get

$$\begin{aligned}\|\nabla\mathcal{I}_h(\eta_k^T w)\|_{L_2(\omega_{k+1}(T)\setminus\omega_{k-3}(T))} &\leq C\|\nabla(\eta_k^T w)\|_{L_2(\omega_{k+1}(T)\setminus\omega_{k-3}(T))}\\ &\leq C\|w\nabla\eta_k^T\|_{L_2(\omega_k(T)\setminus\omega_{k-1}(T))}+C\|\eta_k^T\nabla w\|_{L_2(\omega_{k+1}(T)\setminus\omega_{k-1}(T))}\\ &\leq CH^{-1}\|w-I_Hw\|_{L_2(\omega_k(T)\setminus\omega_{k-1}(T))}+C\|\nabla w\|_{L_2(\omega_{k+1}(T)\setminus\omega_{k-1}(T))}\\ &\leq C\|\nabla w\|_{L_2(\omega_{k+1}(T)\setminus\omega_{k-2}(T))},\end{aligned}$$

where we also have utilized the bounded support of the cut-off function and the bound of I_H in (3.5). This completes the bound (4.8). The bounds in (4.9) and (4.10) follow similarly. □

LEMMA 4.3. *For the Ritz projection* (3.8) *there exist* $\theta\in(0,1)$, *such that*

$$\|\nabla R_{\mathrm{f}}^T v\|_{L_2(\Omega\setminus\omega_k(T))}\leq\theta^k\|\nabla R_{\mathrm{f}}^T v\|_{L_2(\Omega)},\quad v\in V_h, \tag{4.11}$$

where θ *depends on* ρ *and the contrast* β/α, *but not on* k, T, h, H, *or the variations of* A.

PROOF. Fix an element $T\in\mathcal{T}_H$ and let η_k^T be a cut-off function as in (4.6), and define z as in Lemma 4.2 with $w=R_{\mathrm{f}}^T v$ such that

$$z:=(I-I_H)\mathcal{I}_h(\eta_k^T R_{\mathrm{f}}^T v)\in V_{\mathrm{f}}(\Omega\setminus\omega_{k-2}(T)). \tag{4.12}$$

Since $\eta_k^T=1$ on $\Omega\setminus\omega_k(T)$, we have the identity $\mathcal{I}_h\eta_{k,l}^T R_{\mathrm{f}}^T v=R_{\mathrm{f}}^T v$ on $\Omega\setminus\omega_k(T)$. Using this and the bounds (2.7) for $\mathcal{B}(\cdot,\cdot)$ we get

$$\begin{aligned}\|\nabla R_{\mathrm{f}}^T v\|^2_{L_2(\Omega\setminus\omega_k(T))}=\|\nabla(I-I_H)R_{\mathrm{f}}^T v\|^2_{L_2(\Omega\setminus\omega_k(T))}&\leq\|\nabla z\|^2_{L_2(\Omega)}\\ &\leq C_{\mathrm{ko}}^2\alpha^{-1}\mathcal{B}(z,z).\end{aligned} \tag{4.13}$$

Now, due to (4.12) and (4.1), the following equality holds

$$\mathcal{B}(R_{\mathrm{f}}^T v,z)=\mathcal{B}(v,z)_T=0,$$

since z does not have support on the element T. Using this and the fact that $\operatorname{supp}(z-R_{\mathrm{f}}^T v)\cap\operatorname{supp}z\subseteq\omega_k(T)\setminus\omega_{k-2}(T)$ we have

(4.14)

$$\begin{aligned}\mathcal{B}(z,z)=\mathcal{B}(z-R_{\mathrm{f}}^T v,z)&=\int_{\omega_k(T)\setminus\omega_{k-2}(T)}A\varepsilon(z-R_{\mathrm{f}}^T v):\varepsilon(z)\,\mathrm{d}x\\ &\leq\beta\|\nabla(z-R_{\mathrm{f}}^T v)\|_{L_2(\omega_k(T)\setminus\omega_{k-2}(T))}\|\nabla z\|_{L_2(\omega_k(T)\setminus\omega_{k-2}(T))}\\ &\leq\beta\|\nabla(z-R_{\mathrm{f}}^T v)\|_{L_2(\omega_k(T)\setminus\omega_{k-2}(T))}(\|\nabla(z-R_{\mathrm{f}}^T v)\|_{L_2(\omega_k(T)\setminus\omega_{k-2}(T))}\\ &\qquad+\|\nabla R_{\mathrm{f}}^T v\|_{L_2(\omega_k(T)\setminus\omega_{k-2}(T))})\\ &\overset{(4.8)}{\leq}C_{I,\eta}(C_{I,\eta}+1)\beta\|\nabla R_{\mathrm{f}}^T v\|^2_{L_2(\omega_{k+1}(T)\setminus\omega_{k-3}(T))},\end{aligned}$$

Combining (4.13) and (4.14) we have

$$\begin{aligned}\|\nabla R_{\mathrm{f}}^T v\|^2_{L_2(\Omega\setminus\omega_k(T))} &\leq C'\|\nabla R_{\mathrm{f}}^T v\|^2_{L_2(\omega_{k+1}(T)\setminus\omega_{k-3}(T))}\\ &\leq C'(\|\nabla R_{\mathrm{f}}^T v\|^2_{L_2(\Omega\setminus\omega_{k-3}(T))} - \|\nabla R_{\mathrm{f}}^T v\|^2_{L_2(\Omega\setminus\omega_{k+1}(T))}),\end{aligned}$$

where $C' = C_{\mathrm{ko}}^2 C_{I,\eta}(C_{I,\eta}+1)\beta/\alpha$. Thus

$$\|\nabla R_{\mathrm{f}}^T v\|^2_{L_2(\Omega\setminus\omega_{k+1}(T))} \leq \frac{C'}{1+C'}\|\nabla R_{\mathrm{f}}^T v\|^2_{L_2(\Omega\setminus\omega_{k-3}(T))}.$$

An iterative application of this result and relabeling $k+1 \to k$ yields (4.11), with $\theta = ((\frac{C'}{1+C'})^{1/4})^{1/2} < 1$. □

Lemma 4.4. *For the Ritz projections* (4.1) *and* (4.2) *we have the bound*

$$\|\sum_{T\in\mathcal{T}_H} \nabla(R_{\mathrm{f}}^T v - R_{\mathrm{f},k}^T v)\|_{L_2(\Omega)} \leq C k^{d/2}\theta^k \frac{\beta}{\alpha}\Big(\sum_{T\in\mathcal{T}_H}\|\nabla R_{\mathrm{f}}^T v\|^2_{L_2(\Omega)}\Big)^{1/2}, \quad v\in V_h,$$

with θ as in Lemma 4.3 and C depends on $C_{\mathrm{ko}}, C'_{I,\eta}$, and $C''_{I,\eta}$.

Proof. Define $e_{\mathrm{f}} := \sum_{T\in\mathcal{T}_H} R_{\mathrm{f}}^T v - R_{\mathrm{f},k}^T v$ and let η_{k+2}^T be the cut-off function as defined in (4.6). Since $e_{\mathrm{f}} \in V_{\mathrm{f}}$, we define $z_e^T := (I - I_H)\mathcal{I}_h(\eta_{k+2}^T e_{\mathrm{f}})$ as in Lemma 4.2 and note that $\operatorname{supp} z_e^T \subseteq \Omega\setminus\omega_k(T)$. Thus, due to the fact that $\operatorname{supp} R_{\mathrm{f},k}^T v \cap \operatorname{supp} z_e^T = \emptyset$ and (4.1), we have

$$\mathcal{B}(R_{\mathrm{f}}^T v - R_{\mathrm{f},k}^T v, z_e^T) = \mathcal{B}(R_{\mathrm{f}}^T v, z_e^T) = \mathcal{B}(v, z_e^T)_T = 0.$$

Using this and the bounds (2.7) we derive

(4.15)
$$\begin{aligned}\|\nabla e_{\mathrm{f}}\|^2_{L_2(\Omega)} &\leq C_{\mathrm{ko}}\alpha^{-1}\mathcal{B}(e_{\mathrm{f}}, e_{\mathrm{f}}) = C_{\mathrm{ko}}\alpha^{-1}\sum_{T\in\mathcal{T}_H}\mathcal{B}(R_{\mathrm{f}}^T v - R_{\mathrm{f},k}^T v, e_{\mathrm{f}})\\ &= C_{\mathrm{ko}}\alpha^{-1}\sum_{T\in\mathcal{T}_H}\mathcal{B}(R_{\mathrm{f}}^T v - R_{\mathrm{f},k}^T v, e_{\mathrm{f}} - z_e^T).\\ &\leq C_{\mathrm{ko}}\sqrt{\beta}\alpha^{-1}\sum_{T\in\mathcal{T}_H}\|R_{\mathrm{f}}^T v - R_{\mathrm{f},k}^T v\|_{\mathcal{B}(\Omega)}\|\nabla(e_{\mathrm{f}} - z_e^T)\|_{L_2(\omega_{k+2}(T))}.\end{aligned}$$

Now, we use Cauchy-Schwarz inequality for sums and Lemma 4.2 to get

(4.16)
$$\begin{aligned}&\sum_{T\in\mathcal{T}_H}\|R_{\mathrm{f}}^T v - R_{\mathrm{f},k}^T v\|_{\mathcal{B}(\Omega)}\|\nabla(e_{\mathrm{f}} - z_e^T)\|_{L_2(\omega_{k+2}(T))}\\ &\overset{(4.9)}{\leq} C'_{I,\eta}\Big(\sum_{T\in\mathcal{T}_H}\|R_{\mathrm{f}}^T v - R_{\mathrm{f},k}^T v\|^2_{\mathcal{B}(\Omega)}\Big)^{1/2}\Big(\sum_{T\in\mathcal{T}_H}\|\nabla e_{\mathrm{f}}\|^2_{L_2(\omega_{k+3}(T))}\Big)^{1/2}\\ &\leq C'_{I,\eta}C'_\rho k^{d/2}\Big(\sum_{T\in\mathcal{T}_H}\|R_{\mathrm{f}}^T v - R_{\mathrm{f},k}^T v\|^2_{\mathcal{B}(\Omega)}\Big)^{1/2}\|\nabla e_{\mathrm{f}}\|_{L_2(\Omega)}.\end{aligned}$$

In the last inequality we have used the total number of patches overlapping an element T is bounded by $C'_\rho k^{d/2}$, where C'_ρ is a constant depending on the shape regularity of the mesh.

It remains to bound $\|R_{\mathrm{f}}^T v - R_{\mathrm{f},k}^T v\|_{\mathcal{B}(\Omega)}$. For this purpose we define $z_v = (I - I_H)\mathcal{I}_h(\eta_k^T R_{\mathrm{f}}^T v)$ as in Lemma 4.2. Recall that $R_{\mathrm{f}}^T v - z_v \in V_{\mathrm{f}}(\omega_k(T))$. Now, we use Galerkin orthogonality to derive

$$\|R_{\mathrm{f}}^T v - R_{\mathrm{f},k}^T v\|_{\mathcal{B}(\Omega)} \le \|R_{\mathrm{f}}^T v - w\|_{\mathcal{B}(\Omega)}, \quad \forall w \in V_{\mathrm{f}}(\omega_k(T)).$$

Thus, with $w = R_{\mathrm{f}}^T v - z_v \in V_{\mathrm{f}}(\omega_k(T))$ we have

$$\begin{aligned}\|R_{\mathrm{f}}^T v - R_{\mathrm{f},k}^T v\|_{\mathcal{B}(\Omega)} &\le \|z_v\|_{\mathcal{B}(\Omega)} \le \sqrt{\beta}\|\nabla z_v\|_{L_2(\Omega)} \le \sqrt{\beta}\|\nabla z_v\|_{L_2(\Omega\setminus\omega_{k-2})}\\ &\le C''_{I,\eta}\sqrt{\beta}\|\nabla R_{\mathrm{f}}^T v\|_{L_2(\Omega\setminus\omega_{k-3})}.\end{aligned}$$

Using Lemma 4.3 we thus have

$$\|\nabla(R_{\mathrm{f}}^T v - R_{\mathrm{f},k}^T v)\|_{L_2(\Omega)} \le C''_{I,\eta}\sqrt{\beta}\theta^k\|\nabla R_{\mathrm{f}}^T v\|_{L_2(\Omega)}. \tag{4.17}$$

Combining (4.15), (4.16), and (4.17), concludes the proof. □

Remark 4.5. Using the same techniques as in Lemma 4.3 and Lemma 4.4 we can prove (since the right hand side still has support only on a triangle $T \in \mathcal{T}_H$) exponential decay also for the Neumann boundary correctors

$$\|\nabla(\tilde{b}_{\mathrm{f}}^T - \tilde{b}_{\mathrm{f},k}^T)\|_{L_2(\Omega)} \le Ck^{d/2}\theta^k\frac{\beta}{\alpha}\left(\sum_{T\in\mathcal{T}_H}\|\nabla\tilde{b}_{\mathrm{f}}^T\|_{L_2(\Omega)}^2\right)^{1/2}, \quad v \in V_h,$$

with θ as in Lemma 4.3.

We are now ready to prove Theorem 4.1.

Proof of Theorem 4.1. Recall that $u_h = u_{0,h} + g_h$ and $u_{\mathrm{ms},k} = u_{0,\mathrm{ms},k} + \tilde{b}_{\mathrm{f},k} + g_h - R_{\mathrm{f},k}g_h$. Due to (3.1) and (4.4) we have the Galerkin orthogonality

$$\mathcal{B}(u_h - u_{\mathrm{ms},k}, v) = 0, \quad \forall v \in V_{\mathrm{ms},k},$$

which implies

$$\|u_h - u_{\mathrm{ms},k}\|_{\mathcal{B}(\Omega)} \le \|u_h - v - \tilde{b}_{\mathrm{f},k} - g_h + R_{\mathrm{f},k}\|_{\mathcal{B}(\Omega)}, \quad \forall v \in V_{\mathrm{ms},k}.$$

Let $u_{\mathrm{ms}} = u_{0,\mathrm{ms}} + \tilde{b}_{\mathrm{f}} + g_h - R_{\mathrm{f}}g_h$ be the solution to (3.12). Since $u_{0,\mathrm{ms}} \in V_{\mathrm{ms}}$ and $u_{0,\mathrm{ms},k} \in V_{\mathrm{ms},k}$ there exist $v_H, v_{H,k} \in V_H$, such that

$$u_{0,\mathrm{ms}} = v_H - R_{\mathrm{f}}v_H, \quad u_{0,\mathrm{ms},k} = v_{H,k} - R_{\mathrm{f},k}v_{H,k}.$$

Using the Galerkin orthogonality with $v = v_H - R_{\mathrm{f},k}v_H \in V_{\mathrm{ms},k}$ we have

$$\begin{aligned}\|u_h - u_{\mathrm{ms},k}\|_{\mathcal{B}(\Omega)} &\le \|u_h - v_H + R_{\mathrm{f},k}v_H - \tilde{b}_{\mathrm{f},k} - g_h + R_{\mathrm{f},k}g_h\|_{\mathcal{B}(\Omega)}\\ &\le \|u_h - v_H + R_{\mathrm{f}}v_H - \tilde{b}_{\mathrm{f}} - g_h + R_{\mathrm{f}}g_H\|_{\mathcal{B}(\Omega)} + \|R_{\mathrm{f},k}v_H - R_{\mathrm{f}}v_H\|_{\mathcal{B}(\Omega)}\\ &\quad + \|\tilde{b}_{\mathrm{f},k} - \tilde{b}_{\mathrm{f}}\|_{\mathcal{B}(\Omega)} + \|R_{\mathrm{f},k}g_h - R_{\mathrm{f}}g_h\|_{\mathcal{B}(\Omega)}),\end{aligned}$$

From (3.14) in Lemma 3.2 we have

$$\begin{aligned}\|u_h - v_H + R_{\mathrm{f}}v_H - \tilde{b}_{\mathrm{f}} - g_h + R_{\mathrm{f}}g_H\|_{\mathcal{B}(\Omega)} &= \|u_h - u_{\mathrm{ms}}\|_{\mathcal{B}(\Omega)}\\ &\le C_\rho C_{\mathrm{ko}}/\sqrt{\alpha}H\|f\|_{L_2(\Omega)},\end{aligned}$$

and due to Lemma 4.4 and (4.1) we have

$$\begin{aligned}\|R_{\mathrm{f},k}v_H - R_{\mathrm{f}}v_H\|^2_{\mathcal{B}(\Omega)} &\leq \beta\|\nabla(R_{\mathrm{f},k}v_H - R_{\mathrm{f}}v_H)\|^2_{L_2(\Omega)}\\ &\leq C\beta^3/\alpha^2 k^d\theta^{2k}\sum_{T\in\mathcal{T}_H}\|\nabla R^T_{\mathrm{f}}v_H\|^2_{L_2(\Omega)}\\ &\leq C\beta^3/\alpha^2 k^d\theta^{2k}\sum_{T\in\mathcal{T}_H}\|\nabla v_H\|^2_{L_2(T)}\\ &= C\beta^3/\alpha^2 k^d\theta^{2k}\|\nabla v_H\|^2_{L_2(\Omega)}.\end{aligned}$$

Now, since $u_{0,\mathrm{ms}}$ satisfies (3.12) we deduce the stability estimate

$$\begin{aligned}\|u_{0,\mathrm{ms}}\|_{\mathcal{B}(\Omega)} &\leq C(1/\sqrt{\alpha}(\|f\|_{L_2(\Omega)} + \|b\|_{L_2(\Gamma_N)}) + \|\tilde{b}_{\mathrm{f}}\|_{\mathcal{B}(\Omega)} + \|g_h - R_{\mathrm{f}}g_h\|_{\mathcal{B}(\Omega)})\\ &\leq C/\sqrt{\alpha}(\|f\|_{L_2(\Omega)} + \|b\|_{L_2(\Gamma_N)} + \sqrt{\alpha}\|g_h\|_{\mathcal{B}(\Omega)}),\end{aligned}$$

where we have used stability derived from (3.11) and (3.8) in the last inequality. Hence, using that $I_H R_{\mathrm{f}}v_H = 0$ and the stability of I_H (3.6), we get

$$\begin{aligned}\|\nabla v_H\|_{L_2(\Omega)} &= \|\nabla I_H(v_H - R_{\mathrm{f}}v_H)\|_{L_2(\Omega)} \leq C\|\nabla u_{0,\mathrm{ms}}\|_{L_2(\Omega)}\\ &\leq C/\sqrt{\alpha}\|u_{0,\mathrm{ms}}\|_{\mathcal{B}(\Omega)} \leq C/\alpha(\|f\|_{L_2(\Omega)} + \|b\|_{L_2(\Gamma_N)} + \sqrt{\alpha}\|g_h\|_{\mathcal{B}(\Omega)}).\end{aligned}$$

Similarly, we deduce the bounds

$$\begin{aligned}\|\tilde{b}_{\mathrm{f},k} - \tilde{b}_{\mathrm{f}}\|^2_{\mathcal{B}(\Omega)} &\leq C\beta^3/\alpha^2 k^d\theta^{2k}\sum_{\substack{T\in\mathcal{T}_H\\ T\cap\Gamma_N\neq\emptyset}}\|\nabla\tilde{b}^T_{\mathrm{f}}\|^2_{L_2(\Gamma_N)}\\ &\leq C\beta^3/\alpha^4 k^d\theta^{2k}\|b\|^2_{L_2(\Gamma_N)}.\\ \|R_{\mathrm{f},k}g_h - R_{\mathrm{f}}g_h\|^2_{\mathcal{B}(\Omega)} &\leq C\beta^3/\alpha^2 k^d\theta^{2k}\sum_{T\in\mathcal{T}_H}\|\nabla R^T_{\mathrm{f}}g_h\|^2_{L_2(\Omega)}\\ &\leq C\beta^3/\alpha^3 k^d\theta^{2k}\|g_h\|^2_{\mathcal{B}(\Omega)}.\end{aligned}$$

Thus we have

$$\begin{aligned}&\|\nabla(u_h - u_{\mathrm{ms},k})\|_{L_2(\Omega)} \leq C_{\mathrm{ko}}/\sqrt{\alpha}\|u_h - u_{\mathrm{ms},k}\|_{\mathcal{B}(\Omega)}\\ &\quad\leq C/\alpha H\|f\|_{L_2(\Omega)} + C\sqrt{\beta^3/\alpha^5}k^{d/2}\theta^k(\|f\|_{L_2(\Omega)} + \|b\|_{L_2(\Gamma_N)} + \sqrt{\alpha}\|g_h\|_{\mathcal{B}(\Omega)}).\end{aligned}$$

The proof is now complete. □

Remark 4.6. To achieve linear convergence in Theorem 4.1 the size of the patches for the localization should be chosen proportional to $\log H^{-1}$, i.e. $k = c\log(H^{-1})$ for some constant c.

5. Numerical Experiment

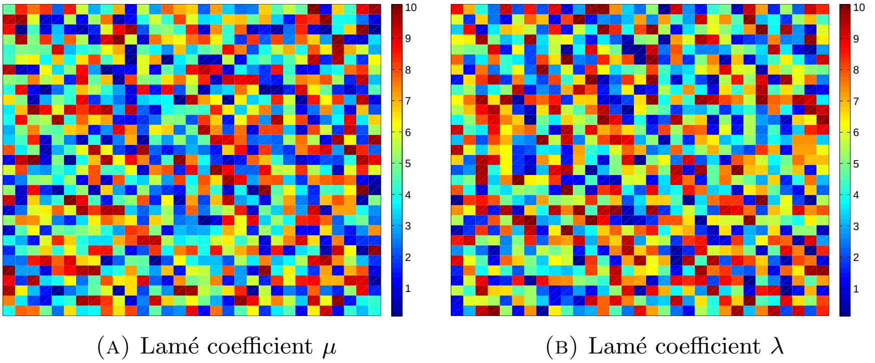
In this section we perform two numerical experiments to test the convergence rate obtained in Theorem 4.1. The first experiment shows that linear convergence is obtained, in the H^1-norm, for a problem with multiscale data. The second experiment shows that the locking effect is reduced for a problem with high value of λ. We refer to [10] for a discussion on how to implement this type of generalized finite elements efficiently.

We consider an isotropic medium, see Remark 2.1, on the unit square in $\mathbb{R}^2$. Recall that the stress tensor in the isotropic case takes the form

$$\sigma(u) = 2\mu\varepsilon(u) + \lambda(\nabla \cdot u)I,$$

where μ and λ are the Lamé coefficients. For simplicity we consider only homogeneous Dirichlet boundary conditions, that is, $\Gamma_D = \partial\Omega$ and $g = 0$. The body forces are set to $f = [1\ 1]^\intercal$.

In the first experiment, we test the convergence on two different setups for the Lamé coefficients, one with multiscale features, and one with constant coefficients $\mu = \lambda = 1$. For the problem with multiscale features we choose μ and λ to be discontinuous on a Cartesian grid of size 2^{-5}. The values at the cells are chosen randomly between 0.1 and 10. The resulting coefficients are shown in Figure 4.



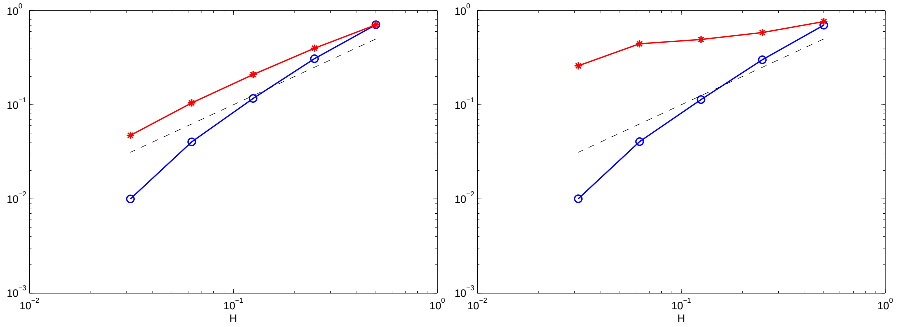
(A) Lamé coefficient μ

(B) Lamé coefficient λ

FIGURE 4. Lamé coefficients with multiscale features.

For the numerical approximations we discretize the domain with a uniform triangulation. The reference solution u_h in (3.1) is computed using a mesh of size $h = \sqrt{2} \cdot 2^{-6}$, which is small enough to resolve the multiscale coefficients in Figure 4. The generalized finite element (GFEM) solution in (4.4) is computed on several meshes of decreasing size, $H = \sqrt{2} \cdot 2^{-1}, ..., \sqrt{2} \cdot 2^{-5}$ with $k = 1, 1, 2, 2, 3$, which corresponds to $k = \lceil 0.8 \log H^{-1} \rceil$. These solutions are compared to the reference solution. For comparison we also compute the classical piecewise linear finite element (P1-FEM) solution on the meshes of size $H = \sqrt{2} \cdot 2^{-1}, ..., \sqrt{2} \cdot 2^{-5}$. The error is computed using the H^1 semi-norm $\|\nabla \cdot \|$ and plotted in Figure 5.

In Figure 5 we see that both methods, as expected, show linear convergence for the problem with constant coefficients. For the problem with multiscale coefficients we clearly see the advantages with the generalized finite element method, which shows linear convergence also in this case, while the classical finite element shows far from optimal convergence.



(A) Constant coefficients $\mu = \lambda = 1$. (B) Multiscale coefficients, see Figure 4.

Figure 5. Relative errors using GFEM (blue $\circ$) and P1-FEM (red $*$) for the linear elasticity problem plotted against the mesh size H. The dashed line is H.

For the second experiment we aim to test the locking effect. We consider a problem from [6]. The domain is set to the unit square $\Omega = [0,1] \times [0,1]$ and $g_D = 0$ on the boundary $\Gamma_D = \partial\Omega$. Furthermore, with $\mu = 1$ and the right hand side $f = [f_1 \ f_2]^\intercal$ chosen as

$$f_1 = \pi^2\Big(4\sin(2\pi y)(-1+2\cos(2\pi x)) - \cos\pi(x+y) + \frac{2}{1+\lambda}\sin(\pi x)\sin(\pi y)\Big),$$
$$f_2 = \pi^2\Big(4\sin(2\pi y)(1-2\cos(2\pi x)) - \cos\pi(x+y) + \frac{2}{1+\lambda}\sin(\pi x)\sin(\pi y)\Big),$$

the exact solution $u = [u_1 \ u_2]^\intercal$ is given by

$$u_1 = \sin(2\pi y)(-1+2\cos(2\pi x)) + \frac{1}{1+\lambda}\sin(\pi x)\sin(\pi y),$$
$$u_2 = \sin(2\pi y)(1-2\cos(2\pi x)) + \frac{1}{1+\lambda}\sin(\pi x)\sin(\pi y).$$

In this experiment we let $\lambda = 10^3$. The discretization of the domain remain the same as in our first example, but the size of the reference mesh is set to $h = \sqrt{2}\cdot 2^{-7}$ which is sufficiently small for u_h to be a relatively good approximation, since $h < 1/\sqrt{\lambda}$. Indeed, using the knowledge of the exact solution we have $\|\nabla(\mathcal{I}_h(u) - u_h)\|_{L_2(\Omega)}/\|\nabla\mathcal{I}_h(u)\|_{L_2(\Omega)} \approx 0.15$, where $\mathcal{I}_h$ is the Lagrangian nodal interpolation onto V_h.

The GFEM and the classical P1-FEM solutions are computed for the values $H = \sqrt{2}\cdot 2^{-1}, ..., \sqrt{2}\cdot 2^{-6}$. The localization parameter is chosen to be $k = 1,1,2,2,3,4$ which corresponds to $k = \lceil 0.8\log H^{-1}\rceil$. The numerical approximations $u_{\mathrm{ms},k}$ and u_H are compared to the reference solution u_h and the error is computed using the H^1-seminorm. The relative errors are plotted in

Figure 6. Clearly, the classical finite element method suffers from locking effects for the coarser mesh sizes. However, the generalized finite element solution shows linear convergence, that is, no locking effect is noted.

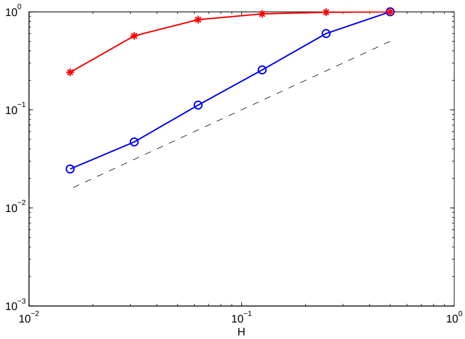


Figure 6. Relative errors for the locking problem using GFEM (blue ∘) and P1-FEM (red $*$) plotted against the mesh size H. The dashed line is H.

References

[1] A. Abdulle: *Analysis of a heterogeneous multiscale FEM for problems in elasticity*, Math. Models Methods Appl. Sci. 16 (2006), no. 4, p. 615–635.

[2] A. Abdulle and P. Henning: *Localized orthogonal decomposition method for the wave equation with a continuum of scales*, to appear in Math. Comp., 2016+.

[3] D. N. Arnold, R. S. Falk and R. Winther: *Mixed finite element methods for linear elasticity with weakly imposed symmetry*, Math. Comp. 76 (2007), no. 260, p. 1699-1723.

[4] I. Babuška and J. E. Osborn: *Generalized finite element methods: their performance and their relation to mixed methods*, SIAM J. Numer. Anal. 20 (1983), no. 3, 1983.

[5] I. Babuška and M. Suri: *Locking effects in the finite-element approximation of elasticity problems*, Numer. Math. 62 (1992), no. 4, p. 439–463.

[6] S. C. Brenner: *A nonconforming mixed multigrid method for the pure displacement problem in planar linear elasticity*, SIAM J. Numer. Anal. 30 (1993), no. 1, p. 116–135.

[7] S. C. Brenner and R. L. Scott: *The mathematical theory of finite element methods,* vol. 15 of *Texts in Applied Mathematics*, Springer, New York, third edition, 2008.

[8] S. C. Brenner and L-Y. Sung: *Linear finite element methods for planar linear elasticity*, Math. Comp. 59 (1992), no. 200, p. 321–338.

[9] F. El Halabi, D. González, A. Chico, and M. Doblaré: FE^2 *multiscale in linear elasticity based on parametrized microscale models using proper generalized decomposition*, Comput. Methods Appl. Mech. Engrg. 257 (2013), p. 183–202.

[10] C. Engwer, P. Henning, A. Målqvist, and D. Peterseim: *Efficient implementation of the Localized Orthogonal Decomposition method*, ArXiv e-print 1602.01658, 2016.

[11] D. Gallistl and D. Peterseim: *Stable multiscale Petrov-Galerkin finite element method for high frequency acoustic scattering*, Comput. Methods Appl. Mech. Engrg. 295 (2015), p. 1–17.

[12] J. M. Gere and B. J. Goodno: *Mechanics of Material*, Cengage Learning, 2008.

[13] P. Henning and A. Målqvist: *Localized orthogonal decomposition techniques for boundary value problems*, SIAM J. Sci. Comput. 36 (2014),

no. 4, p. A1609–A1634.

[14] P. Henning, A. Målqvist and D. Peterseim: *A localized orthogonal decomposition method for semi-linear elliptic problems*, ESAIM Math. Model. Numer. Anal. 48 (2014), no. 5, p. 1331–1349.

[15] P. Henning, A. Målqvist and D. Peterseim: *Two-Level Discretization Techniques for Ground State Computations of Bose-Einstein Condensates*, SIAM J. Numer. Anal. 52 (2014), no. 4, p. 1525–1550.

[16] Henning, Patrick and Peterseim, Daniel: *Oversampling for the Multiscale Finite Element Method*, SIAM Multiscale Model. Simul. 11 (2013), no. 4, p. 1149–1175.

[17] T. J. R. Hughes, G. R. Feijóo, L. Mazzei, and J-B. Quincy: *The variational multiscale method—a paradigm for computational mechanics*, Comput. Methods Appl. Mech. Engrg. 166 (1998), no. 1-2, p. 3–24.

[18] Hughes, T. J. R. and Sangalli, G.: *Variational multiscale analysis: the fine-scale Green's function, projection, optimization, localization, and stabilized methods*, SIAM J. Numer. Anal. 45 (2007), no. 2, p. 539–557.

[19] M. G. Larson and A. Målqvist: *Adaptive variational multiscale methods based on a posteriori error estimation: energy norm estimates for elliptic problems*, Comput. Methods Appl. Mech. Engrg. 196 (2007), no. 21-24, p. 2313–2324.

[20] A. Målqvist: *Multiscale methods for elliptic problems*, Multiscale Model. Simul. 9 (2011), no. 3, p. 1064–1086.

[21] A. Målqvist and A. Persson: *Multiscale techniques for parabolic equations*, Submitted, 2015.

[22] A. Målqvist and D. Peterseim: *Computation of eigenvalues by numerical upscaling*, Numer. Math. 130 (2015), no. 2, p. 337–361.

[23] A. Målqvist and D. Peterseim: *Localization of elliptic multiscale problems*, Math. Comp. 83 (2014), no. 290, p. 2583–2603.

[24] A. Masud and K. Xia: *A variational multiscale method for inelasticity: application to superelasticity in shape memory alloys*, Comput. Methods Appl. Mech. Engrg. 195 (2006), no. 33-36, p. 4512–4531.

[25] A. L. Mazzucato and V. Nistor: *Well-posedness and regularity for the elasticity equation with mixed boundary conditions on polyhedral domains and domains with cracks*, Arch. Ration. Mech. Anal. 195 (2010), no. 1, p. 25–73.

[26] Mott, P. H. and Roland, C. M.: *Limits to Poisson's ratio in isotropic materials*, Physical Review B 80 (2009), no. 13.

[27] Peterseim, D.: *Eliminating the pollution effect in Helmholtz problems by local subscale correction*, ,. Submitted, 2015

[28] D. Peterseim: *Variational Multiscale Stabilization and the Exponential Decay of Fine-scale Correctors*, To appear, 2015/16.

[29] B. Xia and V. H. Hoang: *High-dimensional finite element method for multiscale linear elasticity*, IMA J. Numer. Anal. 35 (2015), no. 3, p. 1277–1314.

Paper III

Axel Målqvist and Anna Persson, *A generalized finite element method for linear thermoelasticity*, Preprint.

A generalized finite element method for linear thermoelasticity

Axel Målqvist[1,2] and Anna Persson[1]

Abstract. We propose and analyze a generalized finite element method designed for linear quasistatic thermoelastic systems with spatial multiscale coefficients. The method is based on the local orthogonal decomposition technique introduced by Målqvist and Peterseim in [18]. We prove convergence of optimal order, independent of the derivatives of the coefficients, in the spatial H^1-norm. The theoretical results are confirmed by numerical examples.

1. Introduction

In many applications the expansion and contraction of a material exposed to temperature changes are of great importance. To model this phenomenon a system consisting of an elasticity equation describing the displacement coupled with an equation for the temperature is used, see, e.g., [6]. The full system consists of a hyperbolic elasticity equation coupled with a parabolic equation for the temperature, see [8] for a comprehensive treatment of this formulation. If the inertia effects are negligible, the hyperbolic term in the elasticity equation can be removed. This leads to an elliptic-parabolic system, often referred to as *quasistatic*. This formulation is discussed in, for instance, [22, 24]. In some settings it is justified to also remove the parabolic term, which leads to an elliptic-elliptic system, see, e.g., [22, 24]. Since the thermoelastic problem is formally equivalent to the system describing poroelasticity, several papers on this equation are also relevant, see, e.g., [5, 25].

In this paper we study the quasistatic case. Existence and uniqueness of a solution to this system are discussed in [22] within the framework of linear degenerate evolution equations in Hilbert spaces. It is also shown that this system is essentially of parabolic type. Existence and uniqueness are also treated in [24] (only two-dimensional problems) and in [23, 21] some results on the thermoelastic contact problem are presented. The classical finite element method for the thermoelastic system is analyzed in [10, 24], where convergence rates of optimal order are derived for problems with solution in H^2 or higher.

When the elastic medium of interest is strongly heterogeneous, like composite materials, the coefficients are highly varying and oscillating. Commonly, such coefficients are said to have *multiscale* features. For these problems classical polynomial finite elements, as in [10, 24], fail to approximate the solution well unless the mesh width resolves the data variations. This is due to the fact that a priori bounds of the error depend on (at least) the spatial H^2-norm of the

[1] Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, SE-412 96 Göteborg, Sweden.

[2] Supported by the Swedish Research Council and the Swedish Foundation for Strategic Research.

solution. Since this norm depends on the derivative of the diffusion coefficient, it is of order ϵ^{-1} if the coefficient oscillates with frequency ϵ^{-1}. To overcome this difficulty, several numerical methods have been proposed, see for instance [4, 3, 11, 18, 14].

In this paper we suggest a generalized finite element method based on the techniques introduced in [18], often referred to as *local orthogonal decomposition*. This method builds on ideas from the variational multiscale method [14, 15], where the solution space is split into a coarse and a fine part. The coarse space is modified such that the basis functions contain information from the diffusion coefficient and have support on small patches. With this approach the basis functions have good approximation properties locally. In [18] the technique is applied to elliptic problems with an arbitrary positive and bounded diffusion coefficient. One of the main advantages is that no assumptions on scale separation or periodicity of the coefficient are needed. Recently, this technique has been applied to several other problems, for instance, semilinear elliptic equations [12], boundary value problems [11], eigenvalue problems [17], linear and semilinear parabolic equations [16], and the linear wave equation [1].

The method we propose in this paper uses generalized finite element spaces similar to those used [18] and [13], together with a correction building on the ideas in [11, 15]. We prove convergence of optimal order that does not depend on the derivatives of the coefficients. We emphasize that by avoiding these derivatives, the a priori bound does not contain any constant of order ϵ^{-1}, although coefficients are highly varying.

In Section 2 we formulate the problem of interest, in Section 3 we first recall the classical finite element method for thermoelasticity and then we define the new generalized finite element method. In Section 4 we perform a localization of the basis functions and in Section 5 we analyze the error. Finally, in Section 6 we present some numerical results.

2. Problem formulation

Let $\Omega \subseteq \mathbb{R}^d$, $d = 2, 3$, be a polygonal/polyhedral domain describing the reference configuration of an elastic body. For a given time $T > 0$ we let $u : [0, T] \times \Omega \to \mathbb{R}^d$ denote the displacement field and $\theta : [0, T] \times \Omega \to \mathbb{R}$ the temperature. To impose Dirichlet and Neumann boundary conditions, we let Γ^u_D and Γ^u_N denote two disjoint segments of the boundary such that $\Gamma := \partial\Omega = \Gamma^u_D \cup \Gamma^u_N$. The segments Γ^θ_D and Γ^θ_N are defined similarly.

We use $(\cdot, \cdot)$ to denote the inner product in $L_2(\Omega)$ and $\|\cdot\|$ for the corresponding norm. Let $H^1(\Omega)$ denote the classical Sobolev space with norm $\|v\|^2_{H^1(\Omega)} = \|v\|^2 + \|\nabla v\|^2$ and let $H^{-1}(\Omega)$ denote the dual space to H^1. Furthermore, we adopt the notation $L_p([0, T]; X)$ for the Bochner space with the norm

$$\|v\|_{L_p([0,T];X)} = \Big(\int_0^T \|v\|_X^p \,\mathrm{dt}\Big)^{1/p}, \quad 1 \leq p < \infty,$$
$$\|v\|_{L_\infty([0,T];X)} = \operatorname*{ess\,sup}_{0\leq t\leq T} \|v\|_X,$$

where X is a Banach space equipped with the norm $\|\cdot\|_X$. The notation $v \in H^1(0,T;X)$ is used to denote $v, \dot{v} \in L_2(0,T;X)$. The dependence on the interval $[0,T]$ and the domain Ω is frequently suppressed and we write, for instance, $L_2(L_2)$ for $L_2([0,T];L_2(\Omega))$. We also define the following subspaces of H^1

$$V^1 := \{v \in (H^1(\Omega))^d : v = 0 \text{ on } \Gamma_D^u\}, \quad V^2 := \{v \in H^1(\Omega) : v = 0 \text{ on } \Gamma_D^\theta\}.$$

Under the assumption that the displacement gradients are small, the (linearized) strain tensor is given by

$$\varepsilon(u) = \frac{1}{2}(\nabla u + \nabla u^{\intercal}).$$

Assuming further that the material is isotropic, Hooke's law gives the (total) stress tensor, see e.g. [21] and the references therein,

$$\bar{\sigma} = 2\mu\varepsilon(u) + \lambda(\nabla\cdot u)I - \alpha\theta I,$$

where I is the d-dimensional identity matrix, α is the thermal expansion coefficient, and μ and λ are the so called Lamé coefficients given by

$$\mu = \frac{E}{2(1+\nu)}, \quad \lambda = \frac{E\nu}{(1+\nu)(1-2\nu)},$$

where E denotes Young's elastic modulus and ν denotes Poisson's ratio. The materials of interest are strongly heterogeneous which implies that α, μ, and λ are rapidly varying in space.

The linear quasistatic thermoelastic problem takes the form

$$-\nabla\cdot(2\mu\varepsilon(u) + \lambda\nabla\cdot uI - \alpha\theta I) = f, \quad \text{in } (0,T]\times\Omega, \tag{2.1}$$
$$\dot{\theta} - \nabla\cdot\kappa\nabla\theta + \alpha\nabla\cdot\dot{u} = g, \quad \text{in } (0,T]\times\Omega, \tag{2.2}$$
$$u = 0, \quad \text{in } (0,T]\times\Gamma_D^u, \tag{2.3}$$
$$\bar{\sigma}\cdot n = 0, \quad \text{in } (0,T]\times\Gamma_N^u. \tag{2.4}$$
$$\theta = 0, \quad \text{on } (0,T]\times\Gamma_D^\theta, \tag{2.5}$$
$$\nabla\theta\cdot n = 0, \quad \text{on } (0,T]\times\Gamma_N^\theta. \tag{2.6}$$
$$\theta(0) = \theta_0, \quad \text{in } \Omega, \tag{2.7}$$

where κ is the heat conductivity parameter, which is assumed to be rapidly varying in space.

REMARK 2.1. For simplicity we have assumed homogeneous boundary data (2.3)-(2.6). However, using techniques similar to the ones used in [11, 13] the analysis in this paper can be extended to non-homogeneous situations.

Assumptions. We make the following assumptions on the data

(A1) $\kappa \in L_\infty(\Omega, \mathbb{R}^{d\times d})$, symmetric,

$$0 < \kappa_1 := \operatorname*{ess\,inf}_{x\in\Omega} \inf_{v\in\mathbb{R}^d\setminus\{0\}} \frac{\kappa(x)v\cdot v}{v\cdot v}, \quad \infty > \kappa_2 := \operatorname*{ess\,sup}_{x\in\Omega} \sup_{v\in\mathbb{R}^d\setminus\{0\}} \frac{\kappa(x)v\cdot v}{v\cdot v},$$

(A2) $\mu, \lambda, \alpha \in L_\infty(\Omega, \mathbb{R})$, and

$$0 < \mu_1 := \operatorname*{ess\,inf}_{x\in\Omega} \mu(x) \le \operatorname*{ess\,sup}_{x\in\Omega} \mu(x) =: \mu_2 < \infty.$$

Similarly, the constants $\lambda_1, \lambda_2, \alpha_1$, and α_2 are used to denote the corresponding upper and lower bounds for λ and α.

(A3) $f, \dot{f} \in L_\infty(L_2), \ddot{f} \in L_\infty(H^{-1})$, $g \in L_\infty(L_2)$, $\dot{g} \in L_\infty(H^{-1})$, and $\theta_0 \in V^2$.

To pose a variational form we multiply the equations (2.1) and (2.2) with test functions from V^1 and V^2 and using Green's formula together with the boundary conditions (2.3)-(2.6) we arrive at the following weak formulation [10]. Find $u(t,\cdot) \in V^1$ and $\theta(t,\cdot) \in V^2$, such that,

$$(\sigma(u) : \varepsilon(v_1)) - (\alpha\theta, \nabla\cdot v_1) = (f, v_1), \quad \forall v_1 \in V^1, \tag{2.8}$$

$$(\dot{\theta}, v_2) + (\kappa\nabla\theta, \nabla v_2) + (\alpha\nabla\cdot\dot{u}, v_2) = (g, v_2), \quad \forall v_2 \in V^2, \tag{2.9}$$

and the initial value $\theta(0,\cdot) = \theta_0$ is satisfied. Here we use σ to denote the effective stress tensor $\sigma(u) := 2\mu\varepsilon(u) + \lambda(\nabla\cdot u)I$ and we use : to denote the Frobenius inner product of matrices. Using Korn's inequality we have the following bounds, see, e.g., [7],

$$c_\sigma \|v_1\|_{H^1}^2 \le (\sigma(v_1) : \epsilon(v_1)) \le C_\sigma \|v_1\|_{H^1}^2, \quad \forall v_1 \in V^1 \tag{2.10}$$

where c_σ (resp. C_σ) depends on μ_1 (resp. μ_2 and λ_2). Similarly, there are constants c_κ (resp. C_κ) depending on the bound κ_1 (resp. κ_2) such that

$$c_\kappa \|v_2\|_{H^1}^2 \le (\kappa\nabla v_2, \nabla v_2) \le C_\kappa \|v_2\|_{H^1}^2, \quad \forall v_2 \in V^2. \tag{2.11}$$

Furthermore, we use the following notation for the energy norms induced by the bilinear forms

$$\|v_1\|_\sigma^2 := (\sigma(v_1) : \varepsilon(v_1)), \; v_1 \in V^1, \quad \|v_2\|_\kappa^2 := (\kappa\nabla v_2 \nabla v_2), \; v_2 \in V^2$$

Existence and uniqueness of a solution to (2.8)-(2.9) have been proved in [22, 24]. There are also some papers on the solution to contact problems, see [2, 23].

Theorem 2.2. Assume that (A1)-(A3) hold and that $\partial\Omega$ is sufficiently smooth. Then there exist u and θ such that $u \in L_2(V^1)$, $\nabla\cdot\dot{u} \in L_2(H^{-1})$, $\theta \in L_2(V^2)$, and $\dot{\theta} \in L_2(H^{-1})$ satisfying (2.8)-(2.9) and the initial condition $\theta(0,\cdot) = \theta_0$.

Remark 2.3. We remark that the equations (2.1)-(2.7) also describe a poroelastic system. In this case θ denotes the fluid pressure, κ the permeability and viscosity of the fluid.

3. Numerical approximation

In this section is we first recall some properties of the classical finite element method for (2.8)-(2.9). In subsection 3.2 we propose a new numerical method built on the ideas from [18]. The localization of this method is treated in Section 4.

3.1. Classical finite element. First, we need to define appropriate finite element spaces. For this purpose we let $\{\mathcal{T}_h\}_{h>0}$ be a family of shape regular triangulations of Ω with the mesh size $h_K := \operatorname{diam}(K)$, for $K \in \mathcal{T}_h$. Furthermore, we denote the largest diameter in the triangulation by $h := \max_{K\in\mathcal{T}_h} h_K$. We now define the classical piecewise affine finite element spaces

$$\begin{aligned} V_h^1 &= \{v \in (C(\bar{\Omega}))^d : v = 0 \text{ on } \Gamma_D^u, v|_K \text{ is a polynomial of deg. } \leq 1, \forall K \in \mathcal{T}_h\}, \\ V_h^2 &= \{v \in C(\bar{\Omega}) : v = 0 \text{ on } \Gamma_D^\theta, v|_K \text{ is a polynomial of deg. } \leq 1, \forall K \in \mathcal{T}_h\}. \end{aligned}$$

For the discretization in time we consider, for simplicity, a uniform time step τ such that $t_n = n\tau$ for $n \in \{0, 1, ..., N\}$ and $N\tau = T$.

Remark 3.1. The classical linear elasticity equation can in some cases suffer from locking effects when using continuous piecewise linear polynomials in both spaces (P1-P1 elements). These typically occur if ν is close to 1/2 (Poisson locking) or if the thickness of the domain is very small (shear locking). In the coupled time-dependent problem locking can occur if $\dot{\theta}$ is neglected in (2.2) and P1-P1 elements are used. The locking produces artificial oscillations in the numerical approximation of the temperature (or pressure) for early time steps. However, it shall be noted that in the case when $\dot{\theta}$ is *not* neglected, this locking effect does not occur, see [20]. Thus, we consider a P1-P1 discretization in this paper.

The classical finite element method with a backward Euler scheme in time reads; for $n \in \{1, ..., N\}$ find $u_h^n \in V_h^1$ and $\theta_h^n \in V_h^2$, such that

$$(\sigma(u_h^n) : \varepsilon(v_1)) - (\alpha\theta_h^n, \nabla \cdot v_1) = (f^n, v_1), \quad \forall v_1 \in V_h^1, \tag{3.1}$$

$$(\bar{\partial}_t\theta_h^n, v_2) + (\kappa\nabla\theta_h^n, \nabla v_2) + (\alpha\nabla \cdot \bar{\partial}_t u_h^n, v_2) = (g^n, v_2), \quad \forall v_2 \in V_h^2, \tag{3.2}$$

where $\bar{\partial}_t\theta_h^n := (\theta_h^n - \theta_h^{n-1})/\tau$ and similarly for $\bar{\partial}_t u_h^n$. The right hand sides are evaluated at time t_n, that is, $f^n := f(t_n)$ and $g^n := g(t_n)$. Given initial data u_h^0 and θ_h^0 the system (3.1)-(3.2) is well posed [10]. We assume that $\theta_h^0 \in V_h^1$ is a suitable approximation of θ_0. For u_h^0 we note that $u(0)$ is uniquely determined by (2.8) at $t = 0$, that is, $u(0)$ fulfills the equation

$$(\sigma(u(0)) : \varepsilon(v_1)) - (\alpha\theta^0, \nabla \cdot v_1) = (f^0, v_1), \quad \forall v_1 \in V^1,$$

and we thus define $u_h^0 \in V_h^1$ to be the solution to

$$(\sigma(u_h^0) : \varepsilon(v_1)) - (\alpha\theta_h^0, \nabla \cdot v_1) = (f^0, v_1), \quad \forall v_1 \in V_h^1. \tag{3.3}$$

The following theorem is a consequence of [10, Theorem 3.1]. The convergence rate is optimal for the two first norms. However, it is not optimal for the

L_2-norm $\|\theta^n - \theta_h^n\|$. In [10] this is avoided by using second order continuous piecewise polynomials for the displacement (P2-P1 elements). It is, however, noted that the problem is still stable using P1-P1 elements. In this paper we use P1-P1 elements and derive error bounds in the $L_\infty(H^1)$-norm, of optimal order, for both the displacement and the temperature.

Theorem 3.2. Let (u, θ) be the solution to (2.8)-(2.9) and $\{(u_h^n, \theta_h^n)\}_{n=1}^N$ be the solution to (3.1)-(3.2). Then for $n \in \{1, ..., N\}$

$$\|u^n - u_h^n\|_{H^1} + \Big(\sum_{m=1}^n \tau \|\theta^m - \theta_h^m\|_{H^1}^2\Big)^{1/2} + \|\theta^n - \theta_h^n\| \leq C_{\epsilon^{-1}}(h + \tau),$$

where $C_{\epsilon^{-1}}$ is of order ϵ^{-1} if the material varies on a scale of size ϵ.

Note that the constant involved in this error bound contains derivatives of the coefficients. Hence, convergence only takes place when the mesh size h is sufficiently small ($h < \epsilon$). Throughout this paper, it is assumed that h is small enough and V_h^1 and V_h^2 are referred to as reference spaces for the solution. Similarly, u_h^n and θ_h^n are referred to as reference solutions. In Section 5 this solution is compared with the generalized finite element solution. We emphasize that the generalized finite element solution is computed in spaces of lower dimension and hence not as computationally expensive.

In the following theorem we prove some regularity results for the finite element solution.

Theorem 3.3. Let $\{u_h^n\}_{n=1}^N$ and $\{\theta_h^n\}_{n=1}^N$ be the solution to (3.1)-(3.2). Then the following bound holds

$$\Big(\sum_{j=1}^n \tau \|\bar\partial_t u_h^j\|_{H^1}^2\Big)^{1/2} + \Big(\sum_{j=1}^n \tau \|\bar\partial_t \theta_h^j\|^2\Big)^{1/2} + \|\theta_h^n\|_{H^1} \leq C(\|g\|_{L_\infty(L_2)} + \|\dot f\|_{L_\infty(H^{-1})} + \|\theta_h^0\|_{H^1}) \tag{3.4}$$

If $\theta_h^0 = 0$, then for $n \in \{1, ..., N\}$

$$\|\bar\partial_t u_h^n\|_{H^1} + \|\bar\partial_t \theta_h^n\| + \Big(\sum_{j=1}^n \tau \|\bar\partial_t \theta_h^j\|_{H^1}^2\Big)^{1/2} \leq C\big(\|g\|_{L_\infty(L_2)} + \|\dot g\|_{L_\infty(H^{-1})} + \|\dot f\|_{L_\infty(H^{-1})} + \|\ddot f\|_{L_\infty(H^{-1})}\big). \tag{3.5}$$

If $f = 0$ and $g = 0$, then for $n \in \{1, ..., N\}$

$$\|\bar\partial_t u_h^n\|_{H^1} + \|\bar\partial_t \theta_h^n\| + t_n^{1/2} \|\bar\partial_t \theta_h^n\|_{H^1} \leq C t_n^{-1/2} \|\theta_h^0\|_{H^1}. \tag{3.6}$$

Proof. From (3.1)-(3.2) and the initial data (3.3) we deduce that the following relation must hold for $n \geq 1$

$$(\sigma(\bar\partial_t u_h^n) : \varepsilon(v_1)) - (\alpha \bar\partial_t \theta_h^n, \nabla \cdot v_1) = (\bar\partial_t f^n, v_1), \quad \forall v_1 \in V_h^1, \tag{3.7}$$

$$(\bar\partial_t \theta_h^n, v_2) + (\kappa \nabla \theta_h^n, \nabla v_2) + (\alpha \nabla \cdot \bar\partial_t u_h^n, v_2) = (g^n, v_2), \qquad \forall v_2 \in V_h^2. \tag{3.8}$$

By choosing $v_1 = \bar{\partial}_t u_h^n$ and $v_2 = \bar{\partial}_t \theta_h^n$ and adding the resulting equations we have

$$\frac{c_\sigma}{2}\|\bar{\partial}_t u_h^n\|_{H^1}^2 + \frac{1}{2}\|\bar{\partial}_t \theta_h^n\|^2 + (\kappa \nabla \theta_h^n, \nabla \bar{\partial}_t \theta_h^n) \leq C(\|g^n\|^2 + \|\bar{\partial}_t f^n\|_{H^{-1}}^2). \tag{3.9}$$

Note that the coupling terms cancel. By using Cauchy-Schwarz and Young's inequality we can bound

$$\tau(\kappa \nabla \theta_h^n, \nabla \bar{\partial}_t \theta_h^n) = \|\kappa^{1/2} \nabla \theta_h^n\|^2 - (\kappa \nabla \theta_h^n, \nabla \theta_h^{n-1}) \geq \frac{1}{2}\|\theta_h^n\|_\kappa^2 - \frac{1}{2}\|\theta_h^{n-1}\|_\kappa^2.$$

Multiplying (3.9) by τ, summing over n, and using (2.10) gives

$$\begin{aligned}\sum_{j=1}^n \tau \|\bar{\partial}_t u_h^j\|_{H^1}^2 + \sum_{j=1}^n \tau \|\bar{\partial}_t \theta_h^j\|^2 + \|\theta_h^n\|_{H^1}^2 \leq C \sum_{j=1}^n \tau(\|g^j\|^2 + \|\bar{\partial}_t f^j\|_{H^{-1}}^2) \\ + C\|\theta_h^0\|_{H^1},\end{aligned}$$

which is bounded by the right hand side in (3.4).

For the bound (3.5) we note that the following relation must hold for $n \geq 2$

$$(\sigma(\bar{\partial}_t u_h^n) : \varepsilon(v_1)) - (\alpha \bar{\partial}_t \theta_h^n, \nabla \cdot v_1) = (\bar{\partial}_t f^n, v_1), \quad \forall v_1 \in V_h^1, \tag{3.10}$$

$$(\bar{\partial}_t^2 \theta_h^n, v_2) + (\kappa \nabla \bar{\partial}_t \theta_h^n, \nabla v_2) + (\alpha \nabla \cdot \bar{\partial}_t^2 u_h^n, v_2) = (\bar{\partial}_t g^n, v_2), \quad \forall v_2 \in V_h^2. \tag{3.11}$$

Now choose $v_1 = \bar{\partial}_t^2 u_h^n$ and $v_2 = \bar{\partial}_t \theta_h^n$ and add the resulting equations to get

$$\begin{aligned}(\sigma(\bar{\partial}_t u_h^n) : \varepsilon(\bar{\partial}_t^2 u_h^n)) + (\bar{\partial}_t^2 \theta_h^n, \bar{\partial}_t \theta_h^n) + (\kappa \nabla \bar{\partial}_t \theta_h^n, \nabla \bar{\partial}_t \theta_h^n) \\ = (\bar{\partial}_t f^n, \bar{\partial}_t^2 u_h^n) + (\bar{\partial}_t g^n, \bar{\partial}_t \theta_h^n).\end{aligned}$$

Multiplying by τ and using Cauchy-Schwarz and Young's inequality gives

$$\begin{aligned}\frac{1}{2}\|\bar{\partial}_t u_h^n\|_\sigma^2 + \frac{1}{2}\|\bar{\partial}_t \theta_h^n\|^2 + C\tau \|\bar{\partial}_t \theta_h^n\|_{H^1}^2 \leq \frac{1}{2}\|\bar{\partial}_t \theta_h^{n-1}\|^2 + \frac{1}{2}\|\bar{\partial}_t u_h^{n-1}\|_\sigma^2 \\ + \tau(\bar{\partial}_t f^n, \bar{\partial}_t^2 u_h^n) + C\|\bar{\partial}_t g^n\|_{H^{-1}}^2.\end{aligned}$$

Summing over n and using (2.10) now gives

$$\begin{aligned}\|\bar{\partial}_t u_h^n\|_{H^1}^2 + \|\bar{\partial}_t \theta_h^n\|^2 + \sum_{j=2}^n \tau \|\bar{\partial}_t \theta_h^j\|_{H^1}^2 \leq C\Big(\|\bar{\partial}_t u_h^1\|_{H^1}^2 + \|\bar{\partial}_t \theta_h^1\|^2 \\ + \sum_{j=2}^n \tau\big((\bar{\partial}_t f^j, \bar{\partial}_t^2 u_h^j) + \|\bar{\partial}_t g^j\|_{H^{-1}}^2\big)\Big).\end{aligned}$$

Here we use summation by parts to get

$$\begin{aligned}\sum_{j=2}^n \tau(\bar{\partial}_t f^j, \bar{\partial}_t^2 u_h^j) &= (\bar{\partial}_t f^n, \bar{\partial}_t u_h^n) - (\bar{\partial}_t f^1, \bar{\partial}_t u_h^1) - \sum_{j=2}^n \tau(\bar{\partial}_t^2 f^j, \bar{\partial}_t u_h^{j-1}) \\ &\leq C\Big(\max_{1\leq j\leq n} \|\bar{\partial}_t f^j\|_{H^{-1}} + \sum_{j=2}^n \tau \|\bar{\partial}_t^2 f^j\|_{H^{-1}}\Big) \max_{1\leq j\leq n} \|\bar{\partial}_t u_h^j\|_{H^1},\end{aligned}$$

and $\max_{1\leq j\leq n} \|\bar{\partial}_t u_h^j\|_{H^1}$ can now be kicked to the left hand side.

To estimate $\bar\partial_t\theta_h^1$ and $\bar\partial_t u_h^1$ we choose $v_1 = \bar\partial_t u_h^1$ and $v_2 = \bar\partial_t\theta_h^1$ in (3.7)-(3.8) for $n = 1$. We thus have, since $\theta_h^0 = 0$,

$$\|\bar\partial_t u_h^1\|_{H^1}^2 + \|\bar\partial_t\theta_h^1\|^2 + \frac{1}{\tau}\|\theta_h^1\|_{H^1}^2 \le C(\|\bar\partial_t f^1\|_{H^{-1}}^2 + \|g^1\|^2).$$

The observation that $\frac{1}{\tau}\|\theta_h^1\|_{H^1}^2 = \tau\|\bar\partial_t\theta_h^1\|_{H^1}^2$ completes the bound (3.5).

Now assume $f = 0$ and $g = 0$ and note that the following holds for $n \ge 2$,

$$(\sigma(\bar\partial_t^2 u_h^n) : \varepsilon(v_1)) - (\alpha\bar\partial_t^2\theta_h^n, \nabla\cdot v_1) = 0, \quad \forall v_1 \in V_h^1,$$
$$(\bar\partial_t^2\theta_h^n, v_2) + (\kappa\nabla\bar\partial_t\theta_h^n, \nabla v_2) + (\alpha\nabla\cdot\bar\partial_t^2 u_h^n, v_2) = 0, \quad \forall v_2 \in V_h^2.$$

Choosing $v_1 = \bar\partial_t^2 u_h^n$, $v_2 = \bar\partial_t^2\theta_h^n$ and adding the resulting equations gives

$$(\sigma(\bar\partial_t^2 u_h^n) : \varepsilon(\bar\partial_t^2 u_h^n)) + (\bar\partial_t^2\theta_h^n, \bar\partial_t^2\theta_h^n) + (\kappa\nabla\bar\partial_t\theta_h^n, \nabla\bar\partial_t^2\theta_h^n) = 0,$$

where, again, the coupling terms cancel. The two first terms on the left hand side are positive and can thus be ignored. Multiplying by τ and t_n^2 gives after using Cauchy-Schwarz and Young's inequality

$$t_n^2\|\bar\partial_t\theta_h^n\|_\kappa^2 \le t_{n-1}^2\|\bar\partial_t\theta_h^{n-1}\|_\kappa^2 + (t_n^2 - t_{n-1}^2)\|\bar\partial_t^2\theta_h^{n-1}\|_\kappa^2.$$

Note that $t_n^2 - t_{n-1}^2 \le 3\tau t_{n-1}$, where we use that $t_n \le 2t_{n-1}$ if $n \ge 2$. Summing over n now gives

$$t_n^2\|\bar\partial_t\theta_h^n\|_\kappa^2 \le t_1^2\|\bar\partial_t\theta_h^1\|_\kappa^2 + 3\sum_{j=2}^n \tau t_{j-1}\|\bar\partial_t\theta_h^{j-1}\|_\kappa^2.$$

To bound the last sum we choose $v_1 = \bar\partial_t^2 u_h^n$, $v_2 = \bar\partial_t\theta_h^n$ in (3.10)-(3.11), now with $f = 0$ and $g = 0$. Adding the resulting equations gives

$$(\bar\partial_t^2\theta_h^n, \bar\partial_t\theta_h^n) + (\kappa\nabla\bar\partial_t\theta_h^n, \nabla\bar\partial_t\theta_h^n) + (\sigma(\bar\partial_t u_h^n) : \varepsilon(\bar\partial_t^2 u_h^n)) = 0,$$

Multiplying by τ and t_n gives after using Cauchy-Schwarz inequality

$$\begin{aligned}&\frac{t_n}{2}\|\bar\partial_t u_h^n\|_\sigma^2 + \frac{t_n}{2}\|\bar\partial_t\theta_h^n\|^2 + c_\kappa\tau t_n\|\bar\partial_t\theta_h^n\|_{H^1}^2 \\ &\quad \le \frac{t_{n-1}}{2}\|\bar\partial_t u_h^{n-1}\|_\sigma^2 + \frac{t_{n-1}}{2}\|\bar\partial_t\theta_h^{n-1}\|^2 + \frac{\tau}{2}\|\bar\partial_t u_h^{n-1}\|_\sigma^2 + \frac{\tau}{2}\|\bar\partial_t\theta_h^{n-1}\|^2.\end{aligned}$$

Summing over n and using (2.10) thus gives

$$\begin{aligned}&\frac{c_\sigma t_n}{2}\|\bar\partial_t u_h^n\|_{H^1}^2 + \frac{t_n}{2}\|\bar\partial_t\theta_h^n\|^2 + \sum_{j=2}^n \tau t_j\|\bar\partial_t\theta_h^j\|_{H^1}^2 \\ &\quad \le \frac{C_\sigma t_1}{2}\|\bar\partial_t u_h^1\|_{H^1}^2 + \frac{t_1}{2}\|\bar\partial_t\theta_h^1\|^2 + C\sum_{j=2}^n \tau\big(\|\bar\partial_t u_h^{j-1}\|_{H^1}^2 + \|\bar\partial_t\theta_h^{j-1}\|^2\big).\end{aligned}$$

To bound the last sum in this estimate we choose $v_1 = \bar\partial_t u_h^n$, $v_2 = \bar\partial_t\theta_h^n$ in (3.7)-(3.8) and multiply by τ to get

$$c_\sigma\tau\|\bar\partial_t u_h^n\|_{H^1}^2 + \tau\|\bar\partial_t\theta_h^n\|^2 + \frac{1}{2}\|\theta_h^n\|_\kappa^2 \le \frac{1}{2}\|\theta_h^{n-1}\|_\kappa^2.$$

Summing over n and using (2.11) gives

$$C\sum_{j=1}^{n}\tau\big(\|\bar{\partial}_t\theta_h^j\|^2+\|\bar{\partial}_t u_h^j\|_{H^1}^2\big)+\frac{c_\kappa}{2}\|\theta_h^n\|_{H^1}^2\leq\frac{C_\kappa}{2}\|\theta_h^0\|_{H^1}^2. \tag{3.12}$$

It remains to bound $t_1^2\|\bar{\partial}_t\theta_h^1\|_{H^1}^2$, $t_1\|\bar{\partial}_t\theta_h^1\|^2$, and $t_1\|\bar{\partial}_t u_h^1\|_{H^1}$. For this purpose we recall that $t_1=\tau$ and use (3.12) for $n=1$ to get

$$\begin{aligned}&t_1\|\bar{\partial}_t u_h^1\|_{H^1}+t_1\|\bar{\partial}_t\theta_h^1\|^2+t_1^2\|\bar{\partial}_t\theta_h^1\|_{H^1}^2\\&\qquad\leq C(\tau(\|\bar{\partial}_t u_h^1\|_{H^1}^2+\|\bar{\partial}_t\theta_h^1\|^2)+\|\theta_h^1\|_{H^1}^2+\|\theta_h^0\|_{H^1}^2)\leq C\|\theta^0\|_{H^1}^2.\end{aligned}$$

Finally, we have that

$$t_n\|\bar{\partial}_t u_h^n\|_{H^1}^2+t_n\|\bar{\partial}_t\theta_h^n\|^2\leq C\|\theta^0\|_{H^1}^2,\quad t_n^2\|\bar{\partial}_t\theta_h^n\|_{H^1}^2\leq C\|\theta^0\|_{H^1}^2,$$

and thus (3.6) follows. □

3.2. Generalized finite element. In this section we shall derive a generalized finite element method. First we define V_H^1 and V_H^2 analogously to V_h^1 and V_h^2, but with a larger mesh size $H>h$. In addition, we assume that the family of triangulations $\{\mathcal{T}_H\}_{H>h}$ is quasi-uniform and that $\mathcal{T}_h$ is a refinement of $\mathcal{T}_H$ such that $V_H^1\subseteq V_h^1$ and $V_H^2\subseteq V_h^2$. Furthermore, we use the notation $\mathcal{N}=\mathcal{N}^1\times\mathcal{N}^2$ to denote the free nodes in $V_H^1\times V_H^2$. The aim is now to define a new (multiscale) space with the same dimension as $V_H^1\times V_H^2$, but with better approximation properties. For this purpose we define an interpolation operator $I_H=(I_H^1,I_H^2):V_h^1\times V_h^2\to V_H^1\times V_H^2$ with the property that $I_H\circ I_H=I_H$ and for all $v=(v_1,v_2)\in V_h^1\times V_h^2$

$$H_K^{-1}\|v-I_Hv\|_{L_2(K)}+\|\nabla I_Hv\|_{L_2(K)}\leq C_I\|\nabla v\|_{L_2(\omega_K)},\quad\forall K\in\mathcal{T}_H, \tag{3.13}$$

where

$$\omega_K:=\mathrm{int}\ \{\hat{K}\in\mathcal{T}_H:\hat{K}\cap K\neq\emptyset\}.$$

Since the mesh is assumed to be shape regular, the estimates in (3.13) are also global, i.e.,

$$H^{-1}\|v-I_Hv\|+\|\nabla I_Hv\|\leq C\|\nabla v\|, \tag{3.14}$$

where C is a constant depending on the shape regularity parameter, $\gamma>0$;

$$\gamma:=\max_{K\in\mathcal{T}_H}\gamma_K,\ \text{with}\ \gamma_K:=\frac{\operatorname{diam}B_K}{\operatorname{diam}K},\ \text{for}\ K\in\mathcal{T}_H, \tag{3.15}$$

where B_K is the largest ball contained in K.

One example of an interpolation that satisfies the above assumptions is $I_H^i=E_H^i\circ\Pi_H^i$, $i=1,2$. Here Π_H^i denotes the piecewise L_2-projection onto $P_1(\mathcal{T}_H)$ ($P_1(\mathcal{T}_H)^d$ if $i=1$), the space of functions that are affine on each triangle $K\in\mathcal{T}_H$. Furthermore, E_H^1 is an averaging operator mapping $(P_1(\mathcal{T}_H))^d$ into V_H^1, by (coordinate wise)

$$(E_H^{1,j}(v))(z)=\frac{1}{\operatorname{card}\{K\in\mathcal{T}_H:z\in K\}}\sum_{K\in\mathcal{T}_H:z\in K}v^j|_K(z),\quad 1\leq j\leq d,$$

where $z \in \mathcal{N}^1$. E_H^2 mapping $\mathcal{P}_H^1$ to V_H^2 is defined similarly. For a further discussion on this interpolation and other available options we refer to [19].

Let us now define the kernels of I_H^1 and I_H^2

$$V_{\mathrm{f}}^1 := \{v \in V_h^1 : I_H^1 v = 0\}, \quad V_{\mathrm{f}}^2 := \{v \in V_h^2 : I_H^2 v = 0\}$$

The kernels are fine scale spaces in the sense that they contain all features that are not captured by the (coarse) finite element spaces V_H^1 and V_H^2. Note that the interpolation leads to the splits $V_h^1 = V_H^1 \oplus V_{\mathrm{f}}^1$ and $V_h^2 = V_H^2 \oplus V_{\mathrm{f}}^2$, meaning that any function $v_1 \in V_h^1$ can be uniquely decomposed as $v_1 = v_{1,H} + v_{1,\mathrm{f}}$, with $v_{1,H} \in V_H^1$ and $v_{1,\mathrm{f}} \in V_{\mathrm{f}}^1$, and similarly for $v_2 \in V_h^2$.

Now, we introduce a Ritz projection onto the fine scale spaces. For this we use the bilinear forms associated with the diffusion in (2.8)-(2.9). The projection of interest is thus $R_{\mathrm{f}} : V_h^1 \times V_h^2 \to V_{\mathrm{f}}^1 \times V_{\mathrm{f}}^2$, such that for all $(v_1, v_2) \in V_h^1 \times V_h^2$, $R_{\mathrm{f}}(v_1, v_2) = (R_{\mathrm{f}}^1 v_1, R_{\mathrm{f}}^2 v_2)$ fulfills

$$(\sigma(v_1 - R_{\mathrm{f}}^1 v_1) : \varepsilon(w_1)) = 0, \quad \forall w_1 \in V_{\mathrm{f}}^1, \tag{3.16}$$

$$(\kappa \nabla(v_2 - R_{\mathrm{f}}^2 v_2), \nabla w_2) = 0, \quad \forall w_2 \in V_{\mathrm{f}}^2. \tag{3.17}$$

Note that this is an uncoupled system and R_{f}^1 and R_{f}^2 are classical Ritz projections.

For any $(v_1, v_2) \in V_h^1 \times V_h^2$ we have, due to the splits of the spaces V_h^1 and V_h^2 above, that

$$v_1 - R_{\mathrm{f}}^1 v_1 = v_{1,H} - R_{\mathrm{f}}^1 v_{1,H}, \quad v_2 - R_{\mathrm{f}}^2 v_2 = v_{2,H} - R_{\mathrm{f}}^2 v_{2,H}.$$

Using this we define the multiscale spaces

$$V_{\mathrm{ms}}^1 := \{v - R_{\mathrm{f}}^1 v : v \in V_H^1\}, \quad V_{\mathrm{ms}}^2 := \{v - R_{\mathrm{f}}^2 v : v \in V_H^2\}. \tag{3.18}$$

Clearly $V_{\mathrm{ms}}^1 \times V_{\mathrm{ms}}^2$ has the same dimension as $V_H^1 \times V_H^2$. Indeed, with λ_x^1 denoting the hat function in V_H^1 at node x and λ_y^2 the hat function in V_H^2 at node y, such that

$$V_H^1 \times V_H^2 = \operatorname{span}\{(\lambda_x^1, 0), (0, \lambda_y^2) : (x, y) \in \mathcal{N}\},$$

a basis for $V_{\mathrm{ms}}^1 \times V_{\mathrm{ms}}^2$ is given by

$$\{(\lambda_x^1 - R_{\mathrm{f}}^1 \lambda_x^1, 0), (0, \lambda_y^2 - R_{\mathrm{f}}^2 \lambda_y^2) : (x, y) \in \mathcal{N}\}. \tag{3.19}$$

Finally, we also note that the splits $V_h^1 = V_{\mathrm{ms}}^1 \oplus V_{\mathrm{f}}^1$ and $V_h^2 = V_{\mathrm{ms}}^2 \oplus V_{\mathrm{f}}^2$ hold, which fulfill the following orthogonality relation

$$(\sigma(v_1) : \varepsilon(w_1) = 0, \quad \forall v_1 \in V_{\mathrm{ms}}^1, \, w_1 \in V_{\mathrm{f}}^1, \tag{3.20}$$

$$(\kappa \nabla v_2, \nabla w_2) = 0, \quad \forall v_2 \in V_{\mathrm{ms}}^2, \, w_2 \in V_{\mathrm{f}}^2 \tag{3.21}$$

3.2.1. *Stationary problem.* For the error analysis in Section 5 it is convenient to define the Ritz projection onto the multiscale space using the bilinear form given by the stationary version of (2.8)-(2.9). We thus define $R_{\mathrm{ms}} : V_h^1 \times V_h^2 \to V_{\mathrm{ms}}^1 \times V_{\mathrm{ms}}^2$, such that for all $(v_1, v_2) \in V_h^1 \times V_h^2$, $R_{\mathrm{ms}}(v_1, v_2) = (R_{\mathrm{ms}}^1(v_1, v_2), R_{\mathrm{ms}}^2 v_2)$ fulfills

$$(\sigma(v_1 - R_{\mathrm{ms}}^1(v_1, v_2)) : \varepsilon(w_1)) - (\alpha(v_2 - R_{\mathrm{ms}}^2 v_2), \nabla \cdot w_1) = 0, \quad \forall w_1 \in V_{\mathrm{ms}}^1, \tag{3.22}$$

$$(\kappa \nabla(v_2 - R_{\mathrm{ms}}^2 v_2), \nabla w_2) = 0, \quad \forall w_2 \in V_{\mathrm{ms}}^2. \tag{3.23}$$

Note that we must have $R_{\mathrm{ms}}^2 = I - R_{\mathrm{f}}^2$, but $R_{\mathrm{ms}}^1 \neq I - R_{\mathrm{f}}^1$ in general.

The Ritz projection in (3.22)-(3.23) is upper triangular. Hence, when solving for $R_{\mathrm{ms}}^1(v_1, v_2)$ the term $(\alpha R_{\mathrm{ms}}^2 v_2, \nabla \cdot w_1)$ in (3.23) is known. Since this term has multiscale features and appears on the right hand side, we impose a correction on $R_{\mathrm{ms}}^1(v_1, v_2)$ inspired by the ideas in [11] and [15]. The correction is defined as the element $\tilde{R}_{\mathrm{f}} v_2 \in V_{\mathrm{f}}^1$, which fulfills

$$(\sigma(\tilde{R}_{\mathrm{f}} v_2) : \varepsilon(w_1)) = (\alpha R_{\mathrm{ms}}^2 v_2, \nabla \cdot w_1), \quad \forall w_1 \in V_{\mathrm{f}}^1, \tag{3.24}$$

and we define $\tilde{R}_{\mathrm{ms}}^1(v_1, v_2) = R_{\mathrm{ms}}^1(v_1, v_2) + \tilde{R}_{\mathrm{f}} v_2$.

Note that the Ritz projections are stable in the sense that

$$\|\tilde{R}_{\mathrm{ms}}^1(v_1, v_2)\|_{H^1} \leq C(\|v_1\|_{H^1} + \|v_2\|_{H^1}), \quad \|R_{\mathrm{ms}}^2 v_2\|_{H^1} \leq C\|v_2\|_{H^1}. \tag{3.25}$$

Remark 3.4. The problem to find $\tilde{R}_{\mathrm{f}} v_2$ is posed in the entire fine scale space and is thus computationally expensive to solve. The aim is to localize these computations to smaller patches of coarse elements, see Section 4.

To derive error bounds for this projection we define two operators $\mathcal{A}_1 : V_h^1 \times V_h^2 \to V_h^1$ and $\mathcal{A}_2 : V_h^2 \to V_h^2$ such that for all $(v_1, v_2) \in V_h^1 \times V_h^2$ we have

$$(\mathcal{A}_1(v_1, v_2), w_1) = (\sigma(v_1) : \varepsilon(w_1)) - (\alpha v_2, \nabla \cdot w_1), \quad \forall w_1 \in V_h^1, \tag{3.26}$$

$$(\mathcal{A}_2 v_2, w_2) = (\kappa \nabla v_2, \nabla w_2), \quad \forall w_2 \in V_h^2. \tag{3.27}$$

Lemma 3.5. For all $(v_1, v_2) \in V_h^1 \times V_h^2$ it holds that

$$\begin{aligned} \|v_1 - \tilde{R}_{\mathrm{ms}}^1(v_1, v_2)\|_{H^1} &\leq C(H\|\mathcal{A}_1(v_1, v_2)\| + \|v_2 - R_{\mathrm{ms}}^2 v_2\|) \\ &\leq CH(\|\mathcal{A}_1(v_1, v_2)\| + \|v_2\|_{H^1}), \end{aligned} \tag{3.28}$$

$$\|v_2 - R_{\mathrm{ms}}^2 v_2\|_{H^1} \leq CH\|\mathcal{A}_2 v_2\|. \tag{3.29}$$

Proof. It follows from [18] that (3.29) holds, since (3.23) is an elliptic equation of Poisson type. Using an Aubin-Nitsche duality argument as in, e.g., [16], we can derive the following estimate in the L_2-norm

$$\|v_2 - R_{\mathrm{ms}}^2 v_2\| \leq CH\|v_2 - R_{\mathrm{ms}}^2 v_2\|_{H^1} \leq CH\|v_2\|_{H^1},$$

which proves the second inequality in (3.28).

It remains to bound $\|v_1 - \tilde{R}_{\mathrm{ms}}^1(v_1, v_2)\|_{H^1}$. Recall that any $v \in V_h^1$ can be decomposed as

$$v = v - R_{\mathrm{f}}^1 v + R_{\mathrm{f}}^1 v = (I - R_{\mathrm{f}}^1)v + R_{\mathrm{f}}^1 v,$$

where $(I - R_{\rm f}^1)v \in V_{\rm ms}^1$. Using the orthogonality (3.20) and that $(\sigma(\cdot) : \varepsilon(\cdot))$ is a symmetric bilinear form we get

$$\begin{aligned}(\sigma(\tilde{R}_{\rm ms}^1(v_1, v_2)) : \varepsilon(v)) &= (\sigma(R_{\rm ms}^1(v_1, v_2) + \tilde{R}_{\rm f}v_2) : \varepsilon((I - R_{\rm f}^1)v + R_{\rm f}^1 v)) \\ &= (\sigma(R_{\rm ms}^1(v_1, v_2)) : \varepsilon((I - R_{\rm f}^1)v)) + (\sigma(\tilde{R}_{\rm f}v_2) : \varepsilon(R_{\rm f}^1 v)).\end{aligned}$$

Due to (3.22) and (3.24) we thus have

$$\begin{aligned}&(\sigma(R_{\rm ms}^1(v_1, v_2)) : \varepsilon((I - R_{\rm f}^1)v)) + (\sigma(\tilde{R}_{\rm f}v_2) : \varepsilon(R_{\rm f}^1 v)) \\ &\quad= (\sigma(v_1) : \varepsilon((I - R_{\rm f}^1)v)) - (\alpha(v_2 - R_{\rm ms}^2 v_2), \nabla\cdot(I - R_{\rm f}^1)v) \\ &\qquad+ (\alpha R_{\rm ms}^2 v_2, \nabla\cdot R_{\rm f}^1 v) \\ &\quad= (\mathcal{A}_1(v_1, v_2), (I - R_{\rm f}^1)v) + (\alpha R_{\rm ms}^2 v_2, \nabla\cdot v).\end{aligned}$$

Define $e := v_1 - \tilde{R}_{\rm ms}^1(v_1, v_2)$. Using the above relation together with (3.26) we get the bound

$$\begin{aligned}c_\sigma\|e\|_{H^1}^2 \le (\sigma(e) : \varepsilon(e)) &= (\sigma(v_1) : \varepsilon(e)) - (\mathcal{A}_1(v_1, v_2), (I - R_{\rm f}^1)e) \\ &\quad- (\alpha R_{\rm ms}^2 v_2, \nabla\cdot e) \\ &= (\mathcal{A}_1(v_1, v_2), R_{\rm f}^1 e) + (\alpha(v_2 - R_{\rm ms}^2 v_2), \nabla\cdot e) \\ &\le \|\mathcal{A}_1(v_1, v_2)\|\|R_{\rm f}^1 e\| + C\|v_2 - R_{\rm ms}^2 v_2\|\|e\|_{H^1}\end{aligned}$$

Since $R_{\rm f}^1 e \in V_{\rm f}^1$ we have due to (3.13)

$$\|R_{\rm f}^1 e\| = \|R_{\rm f}^1 e - I_H^1 R_{\rm f}^1 e\| \le CH\|R_{\rm f}^1 e\|_{H^1} \le CH\|e\|_{H^1},$$

where we have used the stability $\|R_{\rm f}^1 v\|_{H^1} \le C\|v\|_{H^1}$ for $v \in V_h^1$. The first inequality in (3.28) now follows. □

Remark 3.6. Without the correction $\tilde{R}_{\rm f}$ the error bound (3.28) would depend on the derivatives of α,

$$\|v_1 - R_{\rm ms}^1(v_1, v_2)\|_{H^1} \le C_{\alpha'}(H\|\mathcal{A}_1(v_1, v_2)\| + \|v_2 - R_{\rm ms}^2 v_2\|),$$

where α' is large if α has multiscale features.

3.2.2. *Time-dependent problem.* A generalized finite element method with a backward Euler discretization in time is now defined by replacing V_h^1 with $V_{\rm ms}^1$ and V_h^2 with $V_{\rm ms}^2$ in (3.1)-(3.2) and adding a correction similar to (3.24). The method thus reads; for $n \in \{1, ..., N\}$ find $\tilde{u}_{\rm ms}^n = u_{\rm ms}^n + u_{\rm f}^n$, with $u_{\rm ms}^n \in V_{\rm ms}^1$, $u_{\rm f}^n \in V_{\rm f}^1$, and $\theta_{\rm ms}^n \in V_{\rm ms}^2$, such that

$$(\sigma(\tilde{u}_{\rm ms}^n) : \varepsilon(v_1)) - (\alpha\theta_{\rm ms}^n, \nabla\cdot v_1) = (f^n, v_1), \quad \forall v_1 \in V_{\rm ms}^1, \tag{3.30}$$

$$(\bar{\partial}_t\theta_{\rm ms}^n, v_2) + (\kappa\nabla\theta_{\rm ms}^n, \nabla v_2) + (\alpha\nabla\cdot\bar{\partial}_t\tilde{u}_{\rm ms}^n, v_2) = (g^n, v_2), \quad \forall v_2 \in V_{\rm ms}^2, \tag{3.31}$$

$$(\sigma(u_{\rm f}^n) : \varepsilon(w_1)) - (\alpha\theta_{\rm ms}^n, \nabla\cdot w_1) = 0, \qquad \forall w_1 \in V_{\rm f}^1. \tag{3.32}$$

where $\theta_{\rm ms}^0 = R_{\rm ms}^2\theta_h^0$. Furthermore, we define $\tilde{u}_{\rm ms}^0 := u_{\rm ms}^0 + u_{\rm f}^0$, where $u_{\rm f}^0 \in V_{\rm f}^1$ is defined by (3.32) for $n = 0$ and $u_{\rm ms}^0 \in V_{\rm ms}^1$, such that

$$(\sigma(\tilde{u}_{\rm ms}^0) : \varepsilon(v_1)) - (\alpha\theta_{\rm ms}^0, \nabla\cdot v_1) = (f^0, v_1), \quad \forall v_1 \in V_{\rm ms}^1. \tag{3.33}$$

LEMMA 3.7. The problem (3.30)-(3.31) is well-posed.

PROOF. Given u_{ms}^{n-1}, θ_{ms}^{n-1}, and u_{f}^{n-1}, the equations (3.30)-(3.32) yields a square system. Hence, it is sufficient to prove that the solution is unique. Let $v_1 = u_{\text{ms}}^n - u_{\text{ms}}^{n-1}$ in (3.30) and $v_2 = \tau\theta_{\text{ms}}^n$ in (3.31) and add the resulting equations to get

$$\begin{aligned}(\sigma(u_{\text{ms}}^n) : \varepsilon(u_{\text{ms}}^n - u_{\text{ms}}^{n-1})) + (\sigma(u_{\text{f}}^n) : \varepsilon(u_{\text{ms}}^n - u_{\text{ms}}^{n-1})) + \tau(\bar{\partial}_t\theta_{\text{ms}}^n, \theta_{\text{ms}}^n)&\\ + c_\kappa\tau\|\theta_{\text{ms}}^n\|_{H^1}^2 + (\alpha\nabla\cdot(u_{\text{f}}^n - u_{\text{f}}^{n-1}), \theta_{\text{ms},k}^n)&\\ \leq (f^n, u_{\text{ms}}^n - u_{\text{ms}}^{n-1}) + \tau(g^n, \theta_{\text{ms}}^n).&\end{aligned}$$

Using the orthogonality (3.20) and (3.32) this simplifies to

$$\begin{aligned}(\sigma(u_{\text{ms}}^n) : \varepsilon(u_{\text{ms}}^n - u_{\text{ms}}^{n-1})) + \tau(\bar{\partial}_t\theta_{\text{ms}}^n, \theta_{\text{ms}}^n) + c_\kappa\tau\|\theta_{\text{ms}}^n\|_{H^1}^2 + c_\sigma\|u_{\text{f}}^n\|_{H^1}^2&\\ \leq (f^n, u_{\text{ms}}^n - u_{\text{ms}}^{n-1}) + \tau(g^n, \theta_{\text{ms}}^n) + (\sigma(u_{\text{f}}^n) : \varepsilon(u_{\text{f}}^{n-1})).&\end{aligned}$$

Now, using that $(\sigma(\cdot) : \varepsilon(\cdot))$ is a symmetric bilinear form we get the following identity

$$\begin{aligned}(\sigma(v) : \varepsilon(v - w)) = \frac{1}{2}(\sigma(v) : \varepsilon(v)) + \frac{1}{2}(\sigma(v - w) : \varepsilon(v - w))&\\ - \frac{1}{2}(\sigma(w) : \varepsilon(w)),&\end{aligned} \tag{3.34}$$

and using Cauchy-Schwarz and Young's inequality we derive

$$(f^n, u_{\text{ms}}^n - u_{\text{ms}}^{n-1}) \leq C\|f^n\|_{H^{-1}} + \frac{1}{2}(\sigma(u_{\text{ms}}^n - u_{\text{ms}}^{n-1}) : \varepsilon(u_{\text{ms}}^n - u_{\text{ms}}^{n-1})).$$

This, together with the estimate $\tau(\bar{\partial}_t\theta_{\text{ms}}^n, \theta_{\text{ms}}^n) \geq \frac{1}{2}\|\theta_{\text{ms}}^n\|^2 - \frac{1}{2}\|\theta_{\text{ms}}^{n-1}\|^2$ and (2.10), leads to

$$\begin{aligned}\frac{c_\sigma}{2}\|u_{\text{ms}}^n\|_{H^1}^2 + \frac{1}{4}\|\theta_{\text{ms}}^n\|^2 + c_\kappa\tau\|\theta_{\text{ms}}^n\|_{H^1}^2 + \frac{c_\sigma}{2}\|u_{\text{f}}^n\|_{H^1}^2&\\ \leq C(\|f^n\|_{H^{-1}}^2 + \tau\|g^n\|^2 + \|\theta_{\text{ms}}^{n-1}\|^2 + \|u_{\text{ms}}^{n-1}\|_{H^1}^2 + \|u_{\text{f}}^{n-1}\|_{H^1}^2).&\end{aligned}$$

Hence, a unique solution exists. □

4. Localization

In this section we show how to truncate the basis functions, which is motivated by the exponential decay of (3.9). We consider a localization inspired by the one proposed in [11], which is performed by restricting the fine scale space to patches of coarse elements defined by the following; for $K \in \mathcal{T}_H$

$$\begin{aligned}&\omega_0(K) := \text{int } K,\\ &\omega_k(K) := \text{int }\big(\cup\{\hat{K} \in \mathcal{T}_H : \hat{K} \cap \overline{\omega_{k-1}(K)} \neq \emptyset\}\big), \quad k = 1, 2, ...\end{aligned}$$

Now let $V_{\text{f}}^1(\omega_k(K)) := \{v \in V_{\text{f}}^1 : v(z) = 0 \text{ on } (\overline{\Omega}\backslash\Gamma_N^u)\backslash\omega_k(K)\}$ be the restriction of V_{f}^1 to the patch $\omega_k(T)$. We define $V_{\text{f}}^2(\omega_k(K))$ similarly.

The localized fine scale space can now be used to approximate the fine scale part of the basis functions in (3.9), which significantly reduces the computational cost for these problems. Let $(\cdot,\cdot)_\omega$ denote the L_2 inner product over a subdomain

$\omega \subseteq \Omega$ and define the local Ritz projection $R^K_{\mathrm{f},k} : V^1_h \times V^2_h \to V^1_{\mathrm{f}}(\omega_k(K)) \times V^2_{\mathrm{f}}(\omega_k(K))$ such that for all $(v_1, v_2) \in V^1_h \times V^2_h$, $R^K_{\mathrm{f},k}(v_1, v_2) = (R^{K,1}_{\mathrm{f},k} v_1, R^{K,2}_{\mathrm{f},k} v_1)$ fulfills

$$(\sigma(R^{K,1}_{\mathrm{f},k} v_1) : \varepsilon(w_1))_{\omega_k(K)} = (\sigma(v_1) : \varepsilon(w_1))_K, \quad \forall w_1 \in V^1_{\mathrm{f}}(\omega_k(K)), \tag{4.1}$$

$$(\kappa \nabla (R^{K,2}_{\mathrm{f},k} v_2), \nabla w_2)_{\omega_k(K)} = (\kappa \nabla v_2, \nabla w_2)_K, \quad \forall w_2 \in V^2_{\mathrm{f}}(\omega_k(K)). \tag{4.2}$$

Note that if we replace $\omega_k(K)$ with Ω in (4.1)-(4.2) and denote the resulting projection $R^K_{\mathrm{f}}(v_1, v_2) = (R^{K,1}_{\mathrm{f}} v_1, R^{K,2}_{\mathrm{f}} v_2)$, then for all $(v_1, v_2) \in V^1_h \times V^2_h$ we have

$$R_{\mathrm{f}}(v_1, v_2) = \sum_{K \in \mathcal{T}_H} R^K_{\mathrm{f}}(v_1, v_2) = \sum_{K \in \mathcal{T}_H} (R^{K,1}_{\mathrm{f}} v_1, R^{K,2}_{\mathrm{f}} v_2).$$

Motivated by this we now define the localized fine scale projection as

$$R_{\mathrm{f},k}(v_1, v_2) := \sum_{K \in \mathcal{T}_H} R^K_{\mathrm{f},k}(v_1, v_2) = \sum_{K \in \mathcal{T}_H} (R^{K,1}_{\mathrm{f},k} v_1, R^{K,2}_{\mathrm{f},k} v_2), \tag{4.3}$$

and the localized multiscale spaces

$$V^1_{\mathrm{ms},k} := \{v_1 - R^1_{\mathrm{f},k} v_1 : v_1 \in V^1_H\}, \quad V^2_{\mathrm{ms},k} := \{v_2 - R^2_{\mathrm{f},k} v_2 : v_2 \in V^2_H\}, \tag{4.4}$$

with the corresponding localized basis

$$\{(\lambda^1_x - R^1_{\mathrm{f},k} \lambda_x, 0), (0, \lambda^2_y - R^2_{\mathrm{f},k} \lambda_y) : (x, y) \in \mathcal{N}\}. \tag{4.5}$$

4.1. Stationary problem. In this section we define a localized version of the stationary problem (3.22)-(3.23). Let $R_{\mathrm{ms},k} : V^1_h \times V^2_h \to V^1_{\mathrm{ms},k} \times V^2_{\mathrm{ms},k}$, such that for all $(v_1, v_2) \in V^1_h \times V^2_h$, $R_{\mathrm{ms},k}(v_1, v_2) = (R^1_{\mathrm{ms},k}(v_1, v_2), R^2_{\mathrm{ms},k} v_2)$. The method now reads; find

$$\tilde{R}^1_{\mathrm{ms},k}(v_1, v_2) = R^1_{\mathrm{ms},k}(v_1, v_2) + \sum_{K \in \mathcal{T}_H} \tilde{R}^K_{\mathrm{f},k} v_2, \quad \text{where } \tilde{R}^K_{\mathrm{f},k} v_2 \in V^1_{\mathrm{f}}(\omega_k(K)),$$

and $R^2_{\mathrm{ms},k} v_2$ such that

$$\begin{aligned}(\sigma(v_1 - \tilde{R}^1_{\mathrm{ms},k}(v_1, v_2)) &: \varepsilon(w_1)) \\ &- (\alpha(v_2 - R^2_{\mathrm{ms},k} v_2), \nabla \cdot w_1) = 0, \quad \forall w_1 \in V^1_{\mathrm{ms},k},\end{aligned} \tag{4.6}$$

$$(\kappa \nabla (v_2 - R^2_{\mathrm{ms},k} v_2), \nabla w_2) = 0, \quad \forall w_2 \in V^2_{\mathrm{ms},k}. \tag{4.7}$$

$$(\sigma(\tilde{R}^K_{\mathrm{f},k} v_2) : \varepsilon(w)) - (\alpha R^2_{\mathrm{ms},k} v_2, \nabla \cdot w)_K = 0, \quad \forall w \in V^1_{\mathrm{f}}(w_k(K)). \tag{4.8}$$

Note that the Ritz projection is stable in the sense that

$$\|\tilde{R}^1_{\mathrm{ms},k}(v_1, v_2)\|_{H^1} \le C(\|v_1\|_{H^1} + \|v_2\|_{H^1}), \quad \|R^2_{\mathrm{ms},k} v_2\|_{H^1} \le C\|v_2\|_{H^1}. \tag{4.9}$$

The following two lemmas give a bound on the error introduced by the localization.

LEMMA 4.1. For all $(v_1, v_2) \in V_h^1 \times V_h^2$, there exists $\xi \in (0,1)$, such that

$$\|R^1_{\mathrm{f},k} v_1 - R^1_{\mathrm{f}} v_1\|^2_{H^1} \le C k^d \xi^{2k} \sum_{K \in \mathcal{T}_H} \|R^{K,1}_{\mathrm{f}} v_1\|^2_{H^1}, \tag{4.10}$$

$$\|R^2_{\mathrm{f},k} v_2 - R^2_{\mathrm{f}} v_2\|^2_{H^1} \le C k^d \xi^{2k} \sum_{K \in \mathcal{T}_H} \|R^{K,2}_{\mathrm{f}} v_2\|^2_{H^1}, \tag{4.11}$$

$$\|\tilde{R}_{\mathrm{f},k} v_2 - \tilde{R}_{\mathrm{f}} v_2\|^2_{H^1} \le C k^d \xi^{2k} \sum_{K \in \mathcal{T}_H} \|\tilde{R}^{K}_{\mathrm{f}} v_2\|^2_{H^1}. \tag{4.12}$$

The bounds (4.10)-(4.11) are direct results from [13], while (4.12) follows by a slight modification of the right hand side. We omit the proof here.

The next lemma gives a bound for the localized Ritz projection.

LEMMA 4.2. For all $(v_1, v_2) \in V_h^1 \times V_h^2$ there exist $\xi \in (0,1)$ such that

$$\|v_1 - \tilde{R}^1_{\mathrm{ms},k}(v_1, v_2)\|_{H^1} \le C(H + k^{d/2}\xi^k)(\|\mathcal{A}_1(v_1, v_2)\| + \|v_2\|_{H^1}), \tag{4.13}$$

$$\|v_2 - R^2_{\mathrm{ms},k} v_2\|_{H^1} \le C(H + k^{d/2}\xi^k)\|\mathcal{A}_2 v_2\|. \tag{4.14}$$

PROOF. It follows from [11] that (4.14) holds. To prove (4.13) we let $v_H \in V_H^1$ and $v_{H,k} \in V_H^1$ be elements such that

$$R^1_{\mathrm{ms}}(v_1, v_2) = v_H - R^1_{\mathrm{f}} v_H, \quad R^1_{\mathrm{ms},k}(v_1, v_2) = v_{H,k} - R^1_{\mathrm{f},k} v_{H,k}.$$

Define $e := v_1 - \tilde{R}^1_{\mathrm{ms},k}(v_1, v_2)$. From (4.6)-(4.7) we get have the following identity for any $z \in V^1_{\mathrm{ms},k}$

$$\begin{aligned}(\sigma(e) :&\varepsilon(e)) - (\alpha(v_2 - R^2_{\mathrm{ms},k} v_2), \nabla \cdot e) \\ &= (\sigma(e) : \varepsilon(v_1 - z - \tilde{R}_{\mathrm{f},k} v_1)) - (\alpha(v_2 - R^2_{\mathrm{ms},k} v_2), \nabla \cdot (v_1 - z - \tilde{R}_{\mathrm{f},k} v_2)).\end{aligned}$$

Using this with $z = v_H - R^1_{\mathrm{f},k} v_H \in V^1_{\mathrm{ms},k}$ we get

$$\begin{aligned} c_\sigma \|e\|^2_{H^1} \le (\sigma(e) : \varepsilon(e)) = (\sigma(e) &: \varepsilon(v_1 - v_H - R^1_{\mathrm{f},k} v_H - \tilde{R}_{\mathrm{f}} v_1)) \\ &- (\alpha(v_2 - R^2_{\mathrm{ms},k} v_2), \nabla \cdot (v_1 - v_H - R^1_{\mathrm{f},k} v_H - \tilde{R}_{\mathrm{f},k} v_2)) \\ &+ (\alpha(v_2 - R^2_{\mathrm{ms},k} v_2), \nabla \cdot e). \end{aligned}$$

Now, using Cauchy-Schwarz and Young's inequality we get

$$\|e\|^2_{H^1} \le C(\|v_1 - v_H - R^1_{\mathrm{f},k} v_H - \tilde{R}_{\mathrm{f},k} v_2\|^2_{H^1} + \|v_2 - R^2_{\mathrm{ms},k} v_2\|^2),$$

where the last term is bounded in (4.14). For the first term we get

$$\begin{aligned} &\|v_1 - v_H - R^1_{\mathrm{f},k} v_H - \tilde{R}_{\mathrm{f},k} v_2\|_{H^1} \\ &\quad \le \|v_1 - (v_H - R^1_{\mathrm{f}} v_H + \tilde{R}_{\mathrm{f}} v_2)\|_{H^1} + \|R^1_{\mathrm{f}} v_H - R^1_{\mathrm{f},k} v_H\|_{H^1} \\ &\qquad + \|\tilde{R}_{\mathrm{f}} v_2 - \tilde{R}_{\mathrm{f},k} v_2\|^2_{H^1} \\ &\quad \le \|v_1 - \tilde{R}^1_{\mathrm{ms}}(v_1, v_2)\|_{H^1} + \|R^1_{\mathrm{f}} v_H - R^1_{\mathrm{f},k} v_H\|_{H^1} + \|\tilde{R}_{\mathrm{f}} v_2 - \tilde{R}_{\mathrm{f},k} v_2\|_{H^1}, \end{aligned}$$

where the first term on the right hand side is bounded in Lemma 3.5. For the second term we use Lemma 4.1 to get

$$\begin{aligned}\|R_{\mathrm{f}}^1 v_H - R_{\mathrm{f},k}^1 v_H\|_{H^1}^2 &\leq Ck^d\xi^{2k} \sum_{K\in\mathcal{T}_H} \|R_{\mathrm{f}}^{K,1} v_H\|_{H^1}^2 \leq Ck^d\xi^{2k} \sum_{K\in\mathcal{T}_H} \|v_H\|_{H^1(K)}^2 \\ &= Ck^d\xi^{2k}\|v_H\|_{H^1}^2 = Ck^d\xi^{2k}\|I_H(v_H - R_{\mathrm{f}}^1 v_H)\|_{H^1}^2 \\ &= Ck^d\xi^{2k}\|I_H R_{\mathrm{ms}}^1(v_1,v_2)\|_{H^1}^2 \leq Ck^d\xi^{2k}\|R_{\mathrm{ms}}^1(v_1,v_2)\|_{H^1}^2.\end{aligned}$$

We can bound this further by using (3.25) and (3.26), such that

$$\|R_{\mathrm{ms}}^1(v_1,v_2)\|_{H^1} \leq C(\|v_1\|_{H^1} + \|v_2\|_{H^1}) \leq C(\|\mathcal{A}_1(v_1,v_2)\| + \|v_2\|_{H^1}).$$

Similar arguments, using Lemma 4.1 and (4.8), prove

$$\|\tilde{R}_{\mathrm{f}} v_2 - \tilde{R}_{\mathrm{f},k} v_2\|_{H^1} \leq Ck^{d/2}\xi^k\|v_2\|_{H^1},$$

and (4.13) follows. □

Remark 4.3. To preserve linear convergence, the localization parameter k should be chosen such that $k = c\log(H^{-1})$ for some constant c. With this choice of k we get $k^{d/2}\xi^k \sim H$ and we get linear convergence in Lemma 4.2.

We note that the orthogonality relation (3.20) does not hold when V_{ms}^1 is replaced by $V_{\mathrm{ms},k}^1$. However, we have that $V_{\mathrm{ms},k}^1$ and V_{f}^1 are almost orthogonal in the sense that

$$(4.15)\qquad (\sigma(v) : \varepsilon(w)) \leq Ck^{d/2}\xi^k\|v\|_{H^1}\|w\|_{H^1}, \quad \forall v \in V_{\mathrm{ms},k}^1,\, w \in V_{\mathrm{f}}^1.$$

To prove this, note that $v = v_{H,k} - R_{\mathrm{f},k}^1 v_{H,k}$ for some $v_{H,k} \in V_H^1$, and

$$\begin{aligned}(\sigma(v) : \varepsilon(w)) &= (\sigma(v_{H,k} - R_{\mathrm{f}}^1 v_{H,k}) : \varepsilon(w)) + (\sigma(R_{\mathrm{f}}^1 v_{H,k} - R_{\mathrm{f},k}^1 v_{H,k}) : \varepsilon(w)) \\ &= (\sigma(R_{\mathrm{f}}^1 v_{H,k} - R_{\mathrm{f},k}^1 v_{H,k}) : \varepsilon(w)) \\ &\leq C_\sigma\|R_{\mathrm{f}}^1 v_{H,k} - R_{\mathrm{f},k}^1 v_{H,k}\|_{H^1}\|w\|_{H^1},\end{aligned}$$

where we have used that $v_{H,k} - R_{\mathrm{f}}^1 v_{H,k} \in V_{\mathrm{ms}}^1$ and the orthogonality (3.20). Due to Lemma 4.1 we now have

$$\begin{aligned}\|R_{\mathrm{f}}^1 v_{H,k} - R_{\mathrm{f},k}^1 v_{H,k}\|_{H^1}^2 &\leq Ck^d\xi^{2k} \sum_{K\in\mathcal{T}_H} \|R_{\mathrm{f}}^{K,1} v_{H,k}\|_{H^1}^2 \\ &\leq Ck^d\xi^{2k} \sum_{K\in\mathcal{T}_H} \|v_{H,k}\|_{H^1(K)}^2 \\ &= Ck^d\xi^{2k}\|v_{H,k}\|_{H^1}^2 = Ck^d\xi^{2k}\|I_H(v_{H,k} - R_{\mathrm{f},k}^1 v_{H,k})\|_{H^1}^2 \\ &= Ck^d\xi^{2k}\|I_H v\|_{H^1}^2 \leq Ck^d\xi^{2k}\|v\|_{H^1}^2,\end{aligned}$$

and (4.15) follows.

4.2. Time-dependent problem. A localized version of (3.30)-(3.32) is now defined by replacing V^1_{ms} with $V^1_{\mathrm{ms},k}$ and V^2_{ms} with $V^2_{\mathrm{ms},k}$. The method thus reads; for $n \in \{1, ..., N\}$ find

$$\tilde{u}^n_{\mathrm{ms},k} = u^n_{\mathrm{ms},k} + \sum_{K\in\mathcal{T}_H} u^{n,K}_{\mathrm{f},k}, \text{ with } u^n_{\mathrm{ms},k} \in V^1_{\mathrm{ms},k},\ u^{n,K}_{\mathrm{f},k} \in V^1_{\mathrm{f}}(\omega_k(K)),$$

and $\theta^n_{\mathrm{ms},k} \in V^2_{\mathrm{ms},k}$, such that

$$(\sigma(\tilde{u}^n_{\mathrm{ms},k}) : \varepsilon(v_1)) - (\alpha\theta^n_{\mathrm{ms},k}, \nabla\cdot v_1) = (f^n, v_1), \quad \forall v_1 \in V^1_{\mathrm{ms},k}, \tag{4.16}$$

$$\begin{aligned}(\bar{\partial}_t\theta^n_{\mathrm{ms},k}, v_2) &+ (\kappa\nabla\theta^n_{\mathrm{ms},k}, \nabla v_2) \\ &+ (\alpha\nabla\cdot\bar{\partial}_t\tilde{u}^n_{\mathrm{ms},k}, v_2) = (g^n, v_2), \quad \forall v_2 \in V^2_{\mathrm{ms},k},\end{aligned} \tag{4.17}$$

$$(\sigma(u^{n,K}_{\mathrm{f},k}) : \varepsilon(w_1)) - (\alpha\theta^n_{\mathrm{ms},k}, \nabla\cdot w_1)_K = 0, \qquad \forall w_1 \in V^1_{\mathrm{f}}(\omega_k(K)). \tag{4.18}$$

where $\theta^0_{\mathrm{ms},k} = R^2_{\mathrm{ms},k}\theta^0_h$. Furthermore, we define $\tilde{u}^0_{\mathrm{ms},k} = u^0_{\mathrm{ms},k} + \sum_{K\in\mathcal{T}_H} u^{0,K}_{\mathrm{f},k}$, where $u^{0,K}_{\mathrm{f},k} \in V^1_{\mathrm{f}}(\omega_k(K))$ is defined by (4.18) for $n = 0$ and $u^0_{\mathrm{ms},k} \in V^1_{\mathrm{ms}}$ such that

$$(\sigma(\tilde{u}^0_{\mathrm{ms},k}) : \varepsilon(v_1)) - (\alpha\theta^0_{\mathrm{ms},k}, \nabla\cdot v_1) = (f^0, v_1), \quad \forall v_1 \in V^1_{\mathrm{ms},k}. \tag{4.19}$$

We also define $u^n_{\mathrm{f},k} := \sum_{K\in\mathcal{T}_H} u^{n,K}_{\mathrm{f},k}$. Note that for u^n_{f} we have due to (3.32)

$$(\sigma(u^n_{\mathrm{f}}) : \varepsilon(w_1)) - (\alpha\theta^n_{\mathrm{ms}}, \nabla\cdot w_1) = 0, \quad \forall w_1 \in V^1_{\mathrm{f}}.$$

For the localized version $u^n_{\mathrm{f},k}$ this relation is not true. Instead, we prove the following lemma.

Lemma 4.4. For $w_1 \in V^1_{\mathrm{f}}$, it holds that

$$|(\sigma(u^n_{\mathrm{f},k}) : \varepsilon(w_1)) - (\alpha\theta^n_{\mathrm{ms},k}, \nabla\cdot w_1)| \leq Ck^{d/2}\xi^k\|\theta^n_{\mathrm{ms},k}\|\|w_1\|_{H^1}.$$

Proof. Note that from (4.18) we have

$$(\sigma(u^{n,K}_{\mathrm{f},k}) : \varepsilon(w_1)) - (\alpha\theta^n_{\mathrm{ms},k}, \nabla\cdot w_1)_K = 0, \quad \forall w_1 \in V^1_{\mathrm{f}}(\omega_k(K)). \tag{4.20}$$

This equation can be viewed as the localization of the following problem. Find $z^n_{\mathrm{f}} \in V^1_{\mathrm{f}}$, such that

$$(\sigma(z^n_{\mathrm{f}}) : \varepsilon(w_1)) - (\alpha\theta^n_{\mathrm{ms},k}, \nabla\cdot w_1) = 0, \quad \forall w_1 \in V^1_{\mathrm{f}}. \tag{4.21}$$

Now, [13, Lemma 4.4] gives the bound

$$\|z^n_{\mathrm{f}} - u^n_{\mathrm{f},k}\|^2_{H^1} \leq Ck^d\xi^{2k}\sum_{K\in\mathcal{T}_H}\|z^{n,K}_{\mathrm{f}}\|^2_{H^1}$$

where $z^n_{\mathrm{f}} = \sum_{K\in\mathcal{T}_H} z^{n,K}_{\mathrm{f}}$ such that

$$(\sigma(z^{n,K}_{\mathrm{f}}) : \varepsilon(w_1)) - (\alpha\theta^n_{\mathrm{ms},k}, \nabla\cdot w_1)_K = 0, \quad \forall w_1 \in V^1_{\mathrm{f}}.$$

Using this we derive the bound

$$\begin{aligned}(4.22)\quad \|z_{\mathrm{f}}^n - u_{\mathrm{f},k}^n\|_{H^1}^2 &\leq Ck^d\xi^{2k}\sum_{K\in\mathcal{T}_H}\|z_{\mathrm{f}}^{n,K}\|_{H^1}^2 \leq Ck^d\xi^{2k}\sum_{K\in\mathcal{T}_H}\|\theta_{\mathrm{ms},k}^n\|_{L_2(K)}^2\\ &= Ck^d\xi^{2k}\|\theta_{\mathrm{ms},k}^n\|^2.\end{aligned}$$

Now, to prove the lemma we use (4.21) and Cauchy-Schwarz inequality to get

$$\begin{aligned}|(\sigma(u_{\mathrm{f},k}^n):\varepsilon(w_1)) - (\alpha\theta_{\mathrm{ms},k}^n, \nabla\cdot w_1)| &= |(\sigma(u_{\mathrm{f},k}^n - z_{\mathrm{f}}^n):\varepsilon(w_1))|\\ &\leq C_\sigma\|u_{\mathrm{f},k}^n - z_{\mathrm{f}}^n\|_{H^1}\|w_1\|_{H^1}.\end{aligned}$$

Applying (4.22) finishes the proof.

□

The proof can be modified slightly to show the following bound

$$(4.23)\quad |(\sigma(\bar{\partial}_t u_{\mathrm{f},k}^n):\varepsilon(w_1)) - (\alpha\bar{\partial}_t\theta_{\mathrm{ms},k}^n, \nabla\cdot w_1)| \leq Ck^{d/2}\xi^k\|\bar{\partial}_t\theta_{\mathrm{ms},k}^n\|\|w_1\|_{H^1}.$$

Also note that it follows, by choosing $w_1 = u_{\mathrm{f},k}^n$ and $w_1 = \bar{\partial}_t u_{\mathrm{f},k}^n$ respectively, that

$$(4.24)\qquad \|u_{\mathrm{f},k}^n\|_{H^1} \leq C\|\theta_{\mathrm{ms},k}^n\|, \quad \|\bar{\partial}_t u_{\mathrm{f},k}^n\|_{H^1} \leq C\|\bar{\partial}_t\theta_{\mathrm{ms},k}^n\|.$$

To prove that (4.16)-(4.18) is well posed, we need the following condition on the size of H.

Assumptions. We make the following assumption on the size of H.

(A4) $H \leq \min\left(\frac{1}{4C_{\mathrm{co}}}, \frac{c_\sigma}{(C_{\mathrm{co}}+C_{\mathrm{ort}})}\right)$, where C_{co} is the constant in Lemma 4.4 and C_{ort} is the constant in the almost orthogonal property (4.15).

Lemma 4.5. Assuming (A4) the problem (4.16)-(4.18) is well-posed.

Proof. This proof is similar the proof of Lemma 3.7, but we need to account for the lack of orthogonality and the fact that (3.32) is not satisfied.

Given $u_{\mathrm{ms},k}^{n-1}$, $\theta_{\mathrm{ms},k}^{n-1}$, and $u_{\mathrm{f},k}^{n-1} = \sum_K u_{\mathrm{f},k}^{n-1,K}$, the equations (4.16)-(4.18) yields a square system, so it is sufficient to prove that the solution is unique. Choosing $v_1 = u_{\mathrm{ms},k}^n - u_{\mathrm{ms},k}^{n-1}$ in (4.16) and $v_2 = \tau\theta_{\mathrm{ms},k}^n$ in (4.17) and adding the resulting equations we get

$$\begin{aligned}(\sigma(u_{\mathrm{ms},k}^n):\varepsilon(u_{\mathrm{ms},k}^n - u_{\mathrm{ms},k}^{n-1})) + (\sigma(u_{\mathrm{f},k}^n)&:\varepsilon(u_{\mathrm{ms},k}^n - u_{\mathrm{ms},k}^{n-1})) + \tau(\bar{\partial}_t\theta_{\mathrm{ms},k}^n, \theta_{\mathrm{ms},k}^n)\\ &+ c_\kappa\tau\|\theta_{\mathrm{ms},k}^n\|_{H^1}^2 + (\alpha\nabla\cdot(u_{\mathrm{f},k}^n - u_{\mathrm{f},k}^{n-1}), \theta_{\mathrm{ms},k}^n)\\ &\leq (f^n, u_{\mathrm{ms},k}^n - u_{\mathrm{ms},k}^{n-1}) + \tau(g^n, \theta_{\mathrm{ms},k}^n).\end{aligned}$$

Now, using (3.34) and

$$(f^n, u_{\mathrm{ms},k}^n - u_{\mathrm{ms},k}^{n-1}) \leq C\|f^n\|_{H^{-1}} + \frac{1}{2}(\sigma(u_{\mathrm{ms},k}^n - u_{\mathrm{ms},k}^{n-1}):\epsilon(u_{\mathrm{ms},k}^n - u_{\mathrm{ms},k}^{n-1})).$$

together with the estimate $\tau(\bar{\partial}_t\theta^n_{\mathrm{ms},k}, \theta^n_{\mathrm{ms},k}) \geq \frac{1}{2}\|\theta^n_{\mathrm{ms},k}\|^2 - \frac{1}{2}\|\theta^{n-1}_{\mathrm{ms},k}\|^2$, gives

$$\begin{aligned}\frac{c_\sigma}{2}\|u^n_{\mathrm{ms},k}\|^2_{H^1} + \frac{1}{4}\|\theta^n_{\mathrm{ms},k}\|^2 + c_\kappa\tau\|\theta^n_{\mathrm{ms},k}\|^2_{H^1} &+ (\sigma(u^n_{\mathrm{f},k}) : \epsilon(u^n_{\mathrm{ms},k})) \\ &+ (\alpha\nabla\cdot u^n_{\mathrm{f},k}, \theta^n_{\mathrm{ms},k}) \\ &\leq C\|f^n\|^2_{H^{-1}} + \frac{\tau}{2}\|g^n\|^2 + \frac{C_\sigma}{2}\|u^{n-1}_{\mathrm{ms},k}\|^2_{H^1} + \frac{1}{2}\|\theta^{n-1}_{\mathrm{ms},k}\|^2 \\ &\quad + (\sigma(\tilde{u}^n_{\mathrm{f},k}) : \varepsilon(u^{n-1}_{\mathrm{ms},k})) + (\alpha\nabla\cdot u^{n-1}_{\mathrm{f},k}, \theta^n_{\mathrm{ms},k}).\end{aligned}$$

Using Lemma 4.4 we have

$$\begin{aligned}(\alpha\nabla\cdot u^n_{\mathrm{f},k}, \theta^n_{\mathrm{ms},k}) &= (\alpha\theta^n_{\mathrm{ms},k}, \nabla\cdot u^n_{\mathrm{f},k}) - (\sigma(u^n_{\mathrm{f},k}) : \epsilon(u^n_{\mathrm{f},k})) + (\sigma(u^n_{\mathrm{f},k}) : \epsilon(u^n_{\mathrm{f},k})) \\ &\geq -|(\alpha\theta^n_{\mathrm{ms},k}, \nabla\cdot u^n_{\mathrm{f},k}) - (\sigma(u^n_{\mathrm{f},k}) : \epsilon(u^n_{\mathrm{f},k}))| + c_\sigma\|u^n_{\mathrm{f},k}\|^2_{H^1} \\ &\geq -C_{\mathrm{co}}k^{d/2}\xi^k\|u^n_{\mathrm{f},k}\|_{H^1}\|\theta^n_{\mathrm{ms},k}\| + c_\sigma\|u^n_{\mathrm{f},k}\|^2_{H^1}.\end{aligned}$$

and the almost orthogonal property (4.15) gives

$$|(\sigma(u^n_{\mathrm{f},k}) : \varepsilon(u^n_{\mathrm{ms},k}))| \geq -C_{\mathrm{ort}}k^{d/2}\xi^k\|u^n_{\mathrm{f},k}\|_{H^1}\|u^n_{\mathrm{ms},k}\|_{H^1}.$$

Now, using that k should be chosen such that linear convergence is obtained, see Remark 4.3, that is $k^{d/2}\xi^k \sim H$, we conclude after using Young's inequality that

$$\begin{aligned}(\frac{c_\sigma}{2} - \frac{C_{\mathrm{ort}}H}{2})\|u^n_{\mathrm{ms},k}\|^2_{H^1} &+ (\frac{1}{8} - \frac{C_{\mathrm{co}}H}{2})\|\theta^n_{\mathrm{ms},k}\|^2 + c_\kappa\|\theta^n_{\mathrm{ms},k}\|^2_{H^1} \\ &+ (c_\sigma - \frac{(C_{\mathrm{co}} + C_{\mathrm{ort}})H}{2})\|u^n_{\mathrm{f},k}\|^2_{H^1} \\ &\leq C(\|f^n\|^2_{H^{-1}} + \tau\|g^n\|^2 + \|u^{n-1}_{\mathrm{ms},k}\|^2_{H^1} + \|\theta^{n-1}_{\mathrm{ms},k}\|^2 + \|u^{n-1}_{\mathrm{f},k}\|^2_{H^1}),\end{aligned}$$

where assumption (A4) guarantees that the coefficients are positive. Hence, a unique solution exists. □

5. Error analysis

In this section we analyze the error of the generalized finite element method. The results are based on assumption (A4). In the analysis we utilize the following property, which is similar to Lemma 4.4.

LEMMA 5.1. Let $\tilde{e}^n_{\mathrm{f},k} := \tilde{R}_{\mathrm{f},k}\theta^n_h - u^n_{\mathrm{f},k}$ and $\eta^n_\theta := R^2_{\mathrm{ms},k}\theta^n_h - \theta^n_{\mathrm{ms},k}$. Then, for $w_1 \in V^1_{\mathrm{f}}$, it holds that

$$|(\sigma(\tilde{e}^n_{\mathrm{f},k}) : \varepsilon(w_1)) - (\alpha\eta^n_\theta, \nabla\cdot w_1)| \leq Ck^{d/2}\xi^k\|\eta^n_\theta\|\|w_1\|_{H^1}.$$

PROOF. The proof is similar to the proof of Lemma 4.4. We omit the details. □

This can be modified slightly to show the following bound

$$|(\sigma(\bar{\partial}_t\tilde{e}^n_{\mathrm{f},k}) : \varepsilon(w_1)) - (\alpha\bar{\partial}_t\eta^n_\theta, \nabla\cdot w_1)| \leq Ck^{d/2}\xi^k\|\bar{\partial}_t\eta^n_\theta\|\|w_1\|_{H^1}. \tag{5.1}$$

Also note that it follows, by choosing $w_1 = \tilde{e}^n_{\mathrm{f},k}$ and $w_1 = \bar{\partial}_t \tilde{e}^n_{\mathrm{f},k}$ respectively, that

$$\|\tilde{e}^n_{\mathrm{f},k}\|_{H^1} \leq C\|\eta^n_\theta\|, \quad \|\bar{\partial}_t \tilde{e}^n_{\mathrm{f},k}\|_{H^1} \leq C\|\bar{\partial}_t \eta^n_\theta\|. \tag{5.2}$$

Theorem 5.2. Assume that (A4) holds. Let $\{u^n_h\}_{n=1}^N$ and $\{\theta^n_h\}_{n=1}^N$ be the solutions to (3.1)-(3.2) and $\{\tilde{u}^n_{\mathrm{ms},k}\}_{n=1}^N$ and $\{\theta^n_{\mathrm{ms},k}\}_{n=1}^N$ the solutions to (4.16)-(4.18). For $n \in \{1, ..., N\}$ we have

$$\begin{aligned}\|u^n_h - \tilde{u}^n_{\mathrm{ms},k}\|_{H^1} + \|\theta^n_h - \theta^n_{\mathrm{ms},k}\|_{H^1} \leq C(H + k^{d/2}\xi^k)\big(&\|g\|_{L_\infty(L_2)} + \|\dot{g}\|_{L_\infty(H^{-1})} \\ &+ \|f\|_{L_\infty(L_2)} + \|\dot{f}\|_{L_\infty(L_2)} + \|\ddot{f}\|_{L_\infty(H^{-1})} \\ &+ t_n^{-1/2}\|\theta^0_h\|_{H^1}\big).\end{aligned}$$

The proof of Theorem 5.2 is based on two lemmas.

Lemma 5.3. Assume that $\theta^0_h = 0$ and (A4) holds. Let $\{u^n_h\}_{n=1}^N$ and $\{\theta^n_h\}_{n=1}^N$ be the solutions to (3.1)-(3.2) and $\{\tilde{u}^n_{\mathrm{ms},k}\}_{n=1}^N$ and $\{\theta^n_{\mathrm{ms},k}\}_{n=1}^N$ the solutions to (4.16)-(4.18). For $n \in \{1, ..., N\}$ we have

$$\begin{aligned}\|u^n_h - \tilde{u}^n_{\mathrm{ms},k}\|_{H^1} + \|\theta^n_h - \theta^n_{\mathrm{ms},k}\|_{H^1} \leq C(H + k^{d/2}\xi^k)\big(&\|g\|_{L_\infty(L_2)} + \|\dot{g}\|_{L_\infty(H^{-1})} \\ &+ \|f\|_{L_\infty(L_2)} + \|\dot{f}\|_{L_\infty(L_2)} + \|\ddot{f}\|_{L_\infty(H^{-1})}\big).\end{aligned}$$

Proof. We divide the error into the terms

$$\begin{aligned} u^n_h - \tilde{u}^n_{\mathrm{ms},k} &= u^n_h - \tilde{R}^1_{\mathrm{ms},k}(u^n_h, \theta^n_h) + \tilde{R}^1_{\mathrm{ms},k}(u^n_h, \theta^n_h) - \tilde{u}^n_{\mathrm{ms},k} =: \tilde{\rho}^n_u + \tilde{\eta}^n_u, \\ \theta^n_h - \theta^n_{\mathrm{ms},k} &= \theta^n_h - R^2_{\mathrm{ms},k}\theta^n_h + R^2_{\mathrm{ms},k}\theta^n_h - \theta^n_{\mathrm{ms},k} =: \rho^n_\theta + \eta^n_\theta. \end{aligned}$$

We also adopt the following notation

$$\tilde{e}^n_{\mathrm{f},k} := \tilde{R}_{\mathrm{f},k}\theta^n_h - u^n_{\mathrm{f},k}, \quad \eta^n_u := \tilde{\eta}^n_u - \tilde{e}^n_{\mathrm{f},k} = R^1_{\mathrm{ms},k}(u^n_h, \theta^n_h) - u^n_{\mathrm{ms},k}.$$

From (3.2) it follows that

$$(\kappa \nabla \theta^n_h, \nabla v_2) = (g^n - \bar{\partial}_t \theta^n_h - \nabla \cdot \bar{\partial}_t u^n_h, v_2), \quad \forall v_2 \in V^2_h,$$

so by Lemma 4.2 we have the bound

$$\|\rho^n_\theta\|_{H^1} \leq C(H + k^{d/2}\xi^k)\|P^2_h g^n - \bar{\partial}_t \theta^n_h - \nabla \cdot \bar{\partial}_t u^n_h\|,$$

where P^2_h denotes the L_2-projection onto V^2_h. Theorem 3.3 now completes this bound. Similarly, (3.1) gives

$$(\sigma(u^n_h) : \varepsilon(v_1)) - (\alpha \theta^n_h, \nabla \cdot v_1) = (f^n, v_1), \quad \forall v_1 \in V^1_h,$$

so, again, by Lemma 4.2 we get

$$\|\tilde{\rho}^n_u\|_{H^1} \leq C(H + k^{d/2}\xi^k)(\|f^n\| + \|\theta^n_h\|_{H^1}),$$

which can be further bounded by using Theorem 3.3. To bound $\tilde{\eta}^n_u$ and η^n_θ we note that for $v_1 \in V^1_{\mathrm{ms},k}$

$$\begin{aligned} (\sigma(\tilde{\eta}^n_u) : \varepsilon(v_1)) - (\alpha \eta^n_\theta, \nabla \cdot v_1) &\\ = (\sigma(\tilde{R}^1_{\mathrm{ms},k}(u^n_h, \theta^n_h)) &: \varepsilon(v_1)) - (\alpha R^2_{\mathrm{ms},k}\theta^n_h, \nabla \cdot v_1) - (f^n, v_1) \\ = (\sigma(u^n_h) &: \varepsilon(v_1)) - (\alpha \theta^n_h, \nabla \cdot v_1) - (f^n, v_1) = 0, \end{aligned} \tag{5.3}$$

where we have used the Ritz projection (4.6), and the equations (3.1) and (4.16). Similarly, for $v_2 \in V^2_{\mathrm{ms},k}$ we have

$$\begin{aligned}
&(\bar{\partial}_t \eta_\theta^n, v_2) + (\kappa \nabla \eta_\theta^n, \nabla v_2) + (\alpha \nabla \cdot \bar{\partial}_t \tilde{\eta}_u^n, v_2) \\
&\quad = (\bar{\partial}_t R^2_{\mathrm{ms},k} \theta_h^n, v_2) + (\kappa \nabla R^2_{\mathrm{ms},k} \theta_h^n, \nabla v_2) + (\alpha \nabla \cdot \bar{\partial}_t \tilde{R}^1_{\mathrm{ms},k}(u_h^n, \theta_h^n), v_2) \\
&\qquad - (g^n, v_2) \\
&\quad = (-\bar{\partial}_t \rho_\theta^n, v_2) + (-\alpha \nabla \cdot \bar{\partial}_t \tilde{\rho}_u^n, v_2)
\end{aligned}$$

For simplicity, we denote $\rho^n := \rho_\theta^n + \alpha \nabla \cdot \tilde{\rho}_u^n$ such that

$$(\bar{\partial}_t \eta_\theta^n, v_2) + (\kappa \nabla \eta_\theta^n, \nabla v_2) + (\alpha \nabla \cdot \bar{\partial}_t \tilde{\eta}_u^n, v_2) = (-\bar{\partial}_t \rho^n, v_2), \quad \forall v_2 \in V^2_{\mathrm{ms},k} \tag{5.4}$$

Now, choose $v_1 = \bar{\partial}_t \eta_u^n$ and $v_2 = \eta_\theta^n$ and add the resulting equations. Note that the coupling terms on the left hand side results in the term $(\alpha \nabla \cdot \bar{\partial}_t \tilde{e}^n_{\mathrm{f},k}, \eta_\theta^n)$. We conclude that

$$(\sigma(\tilde{\eta}_u^n) : \varepsilon(\bar{\partial}_t \eta_u^n)) + (\bar{\partial}_t \eta_\theta^n, \eta_\theta^n) + (\kappa \nabla \eta_\theta^n, \nabla \eta_\theta^n) = (-\bar{\partial}_t \rho^n, \eta_\theta^n) - (\alpha \nabla \cdot \bar{\partial}_t \tilde{e}^n_{\mathrm{f},k}, \eta_\theta^n),$$

and by splitting the first term

$$\begin{aligned}
&(\sigma(\eta_u^n) : \varepsilon(\bar{\partial}_t \eta_u^n)) + (\bar{\partial}_t \eta_\theta^n, \eta_\theta^n) + (\kappa \nabla \eta_\theta^n, \nabla \eta_\theta^n) \\
&\qquad = (-\bar{\partial}_t \rho^n, \eta_\theta^n) - (\sigma(\tilde{e}^n_{\mathrm{f},k}) : \varepsilon(\bar{\partial}_t \eta_u^n)) - (\alpha \nabla \cdot \bar{\partial}_t \tilde{e}^n_{\mathrm{f},k}, \eta_\theta^n).
\end{aligned} \tag{5.5}$$

Using Lemma 5.1 we can bound

$$\begin{aligned}
-(\alpha \nabla \cdot \bar{\partial}_t \tilde{e}^n_{\mathrm{f},k}, \eta_\theta^n) &\leq |(\alpha \nabla \cdot \bar{\partial}_t \tilde{e}^n_{\mathrm{f},k}, \eta_\theta^n) - (\sigma(\tilde{e}^n_{\mathrm{f},k}) : \varepsilon(\bar{\partial}_t \tilde{e}^n_{\mathrm{f},k}))| \\
&\qquad - (\sigma(\tilde{e}^n_{\mathrm{f},k}) : \varepsilon(\bar{\partial}_t \tilde{e}^n_{\mathrm{f},k})) \\
&\leq C k^{d/2} \xi^k \|\bar{\partial}_t \tilde{e}^n_{\mathrm{f},k}\|_{H^1} \|\eta_\theta^n\| - (\sigma(\tilde{e}^n_{\mathrm{f},k}) : \varepsilon(\bar{\partial}_t \tilde{e}^n_{\mathrm{f},k})),
\end{aligned} \tag{5.6}$$

and the almost orthogonal property (4.15) together with (5.2) gives

$$-(\sigma(\tilde{e}^n_{\mathrm{f},k}) : \varepsilon(\bar{\partial}_t \eta_u^n)) \leq C k^{d/2} \xi^k \|\tilde{e}^n_{\mathrm{f},k}\|_{H^1} \|\bar{\partial}_t \eta_u^n\|_{H^1} \leq C k^{d/2} \xi^k \|\eta_\theta^n\| \|\bar{\partial}_t \eta_u^n\|_{H^1}. \tag{5.7}$$

Thus, multiplying (5.5) by τ and using Cauchy-Schwarz and Young's inequality we get

$$\begin{aligned}
C\tau \|\eta_\theta^n\|^2_{H^1} + \frac{1}{2}(\|\eta_u^n\|^2_\sigma + \|\tilde{e}^n_{\mathrm{f},k}\|^2_\sigma - \|\tilde{\eta}_u^{n-1}\|^2_\sigma - \|\tilde{e}^{n-1}_{\mathrm{f},k}\|^2_\sigma) + \frac{1}{2}(\|\eta_\theta^n\|^2 - \|\eta_\theta^{n-1}\|^2) \\
\leq C\tau \|\bar{\partial}_t \rho^n\|^2_{H^{-1}} + C\tau k^{d/2} \xi^k \|\eta_\theta^n\| (\|\bar{\partial}_t \eta_u^n\|_{H^1} + \|\bar{\partial}_t \tilde{e}^n_{\mathrm{f},k}\|_{H^1}),
\end{aligned}$$

where $\|\eta_\theta^n\| \leq C \|\eta_\theta^n\|_{H^1}$ can be kicked to the left hand side. Summing over n gives

$$\begin{aligned}
&C \sum_{j=1}^n \tau \|\eta_\theta^j\|^2_{H^1} + \frac{1}{2}(\|\eta_u^n\|^2_\sigma + \|\tilde{e}^n_{\mathrm{f},k}\|^2_\sigma) + \frac{1}{2}\|\eta_\theta^n\|^2 \\
&\quad \leq \frac{c_\sigma}{2} \|\tilde{\eta}_u^0\|^2_{H^1} + C \sum_{j=1}^n \tau (\|\bar{\partial}_t \rho^j\|^2_{H^{-1}} + k^d \xi^{2k} (\|\bar{\partial}_t \eta_u^j\|^2_{H^1} + \|\bar{\partial}_t \tilde{e}^j_{\mathrm{f},k}\|^2_{H^1})),
\end{aligned}$$

where we have used that $\eta_\theta^0 = 0$. Furthermore, we note that if $\theta_h^0 = 0$, then $\tilde{R}_{\mathrm{f},k}\theta_h^0 = 0$ and $u_{\mathrm{f},k}^0 = 0$. Hence, $e_{\mathrm{f},k}^0 = 0$. From (4.19) and (3.3) we have, if $\theta_h^0 = \theta_{\mathrm{ms},k}^0 = 0$, for $v_1 \in V_{\mathrm{ms},k}^1$,

$$(\sigma(u_{\mathrm{ms},k}^0) : \varepsilon(v_1)) = (f^0, v_1) = (\sigma(u_h^0) : \varepsilon(v_1)) = (\sigma(R_{\mathrm{ms},k}^1(u_h^0, 0)) : \varepsilon(v_1)),$$

so also $\eta_u^0 = 0$.

To bound $\bar{\partial}_t \rho_\theta^j$ and $\alpha\nabla \cdot \bar{\partial}_t \tilde{\rho}_u^j$ we note that due to (3.1) and (3.3), $\bar{\partial}_t u_h^n$ and $\bar{\partial}_t \theta_h^n$ satisfy the equation

$$(\sigma(\bar{\partial}_t u_h^n) : \varepsilon(v_1)) - (\alpha \bar{\partial}_t \theta_h^n, \nabla \cdot v_1) = (\bar{\partial}_t f^n, v_1), \quad \forall v_1 \in V_h^1.$$

Hence, by Lemma 4.2 and the Aubin-Nitsche duality argument we have

$$\begin{aligned} \|\bar{\partial}_t \rho_\theta^j\|_{H^{-1}} \le \|\bar{\partial}_t \rho_\theta^j\| &\le C(H + k^{d/2}\xi^k)\|\bar{\partial}_t \rho_\theta^j\|_{H^1} \\ &\le C(H + k^{d/2}\xi^k)\|\bar{\partial}_t \theta_h^j\|_{H^1}, \end{aligned} \tag{5.8}$$

and for $\bar{\partial}_t \tilde{\rho}_u^j$ we get

$$\begin{aligned} &\|\alpha \nabla \cdot \bar{\partial}_t \tilde{\rho}_u^j\|_{H^{-1}} \\ &\quad \le \alpha_2 \|\nabla \cdot \bar{\partial}_t \tilde{\rho}_u^j\| \le C\|\bar{\partial}_t \tilde{\rho}_u^j\|_{H^1} \le C(H + k^{d/2}\xi^k)(\|\bar{\partial}_t f^j\| + \|\bar{\partial}_t \theta_h^j\|_{H^1}). \end{aligned} \tag{5.9}$$

Thus, using (2.10), we arrive at the following bound

$$\begin{aligned} \sum_{j=1}^n \tau \|\eta_\theta^j\|_{H^1}^2 + \|\eta_u^n\|_{H^1}^2 + \|\tilde{e}_{\mathrm{f},k}^n\|_{H^1}^2 + \|\eta_\theta^n\|^2 \\ \le C(H + k^{d/2}\xi^k)^2 \sum_{j=1}^n \tau\big(\|\bar{\partial}_t \theta_h^j\|_{H^1}^2 + \|\bar{\partial}_t f^j\|^2\big) \\ + Ck^d \xi^{2k} \sum_{j=1}^n \tau(\|\bar{\partial}_t \eta_u^j\|_{H^1}^2 + \|\bar{\partial}_t \tilde{e}_{\mathrm{f},k}^j\|_{H^1}^2), \end{aligned} \tag{5.10}$$

where we apply Theorem 3.3 to the first sum on the right hand side. If we can find an upper bound on $\sum_{j=1}^n \tau(\|\bar{\partial}_t \eta_u^j\|_{H^1}^2 + \|\bar{\partial}_t \tilde{e}_{\mathrm{f},k}^j\|^2)$, then (5.10) gives a bound for $\|\tilde{\eta}_u^n\|_{H^1} \le \|\eta_u^n\|_{H^1} + \|\tilde{e}_{\mathrm{f},k}^n\|_{H^1}$. This is done next, and we bound $\|\eta_\theta^n\|_{H^1}$ at the same time. For this purpose, we choose $v_2 = \bar{\partial}_t \eta_\theta^n$ in (5.4) and note that it follows from (5.3) that

$$(\sigma(\bar{\partial}_t \tilde{\eta}_u^n) : \varepsilon(\bar{\partial}_t \eta_u^n)) - (\alpha \bar{\partial}_t \eta_\theta^n, \nabla \cdot \bar{\partial}_t \eta_u^n) = 0. \tag{5.11}$$

This also holds for $n = 1$ since $\eta_\theta^0 = 0$ and $\tilde{\eta}_u^0 = 0$. Thus, by adding the resulting equations, we have

$$\begin{aligned} c_\sigma \|\bar{\partial}_t \eta_u^n\|_{H^1}^2 &+ \|\bar{\partial}_t \eta_\theta^n\|^2 + (\kappa \nabla \eta_\theta^n, \nabla \bar{\partial}_t \eta_\theta^n) \\ &= (-\bar{\partial}_t \rho^n, \bar{\partial}_t \eta_\theta^n) - (\sigma(\bar{\partial}_t \tilde{e}_{\mathrm{f},k}^n) : \varepsilon(\bar{\partial}_t \eta_u^n)) - (\alpha \nabla \cdot \bar{\partial}_t \tilde{e}_{\mathrm{f},k}^n, \bar{\partial}_t \eta_\theta^n) \\ &\le \|\bar{\partial}_t \rho^n\| \|\bar{\partial}_t \eta_\theta^n\| + C_{\mathrm{ort}} k^{d/2} \xi^k \|\bar{\partial}_t \tilde{e}_{\mathrm{f},k}^n\|_{H^1} \|\bar{\partial}_t \eta_u^n\|_{H^1} \\ &\quad - (\alpha \nabla \cdot \bar{\partial}_t \tilde{e}_{\mathrm{f},k}^n, \bar{\partial}_t \eta_\theta^n) \end{aligned}$$

where we have used (4.15). For the last term we use Lemma 5.1 to achieve

$$-(\alpha\nabla\cdot\bar{\partial}_t\tilde{e}^n_{\mathrm{f},k},\bar{\partial}_t\eta^n_\theta)\leq C_{\mathrm{co}}k^{d/2}\xi^k\|\bar{\partial}_t\tilde{e}^n_{\mathrm{f},k}\|_{H^1}\|\bar{\partial}_t\eta^n_\theta\|-(\sigma(\bar{\partial}_t\tilde{e}^n_{\mathrm{f},k}):\varepsilon(\bar{\partial}_t\tilde{e}^n_{\mathrm{f},k})).$$

Thus, we have

$$\begin{aligned}c_\sigma(\|\bar{\partial}_t\eta^n_u\|^2_{H^1}+\|\bar{\partial}_t\tilde{e}^n_{\mathrm{f},k}\|^2_{H^1})+\|\bar{\partial}_t\eta^n_\theta\|^2+(\kappa\nabla\eta^n_\theta,\nabla\bar{\partial}_t\eta^n_\theta)&\\ \leq\|\bar{\partial}_t\rho^n\|\|\bar{\partial}_t\eta^n_\theta\|+C_{\mathrm{ort}}k^{d/2}\xi^k\|\bar{\partial}_t\tilde{e}^n_{\mathrm{f},k}\|_{H^1}\|\bar{\partial}_t\eta^n_u\|_{H^1}&\\ +C_{\mathrm{co}}k^{d/2}\xi^k\|\bar{\partial}_t\tilde{e}^n_{\mathrm{f},k}\|_{H^1}\|\bar{\partial}_t\eta^n_\theta\|,&\end{aligned}$$

and using Young's inequality we deduce

$$\begin{aligned}(c_\sigma-\frac{C_{\mathrm{ort}}k^{d/2}\xi^k}{2})\|\bar{\partial}_t\eta^n_u\|^2_{H^1}+(c_\sigma-\frac{(C_{\mathrm{ort}}+C_{\mathrm{co}})k^{d/2}\xi^k}{2})\|\bar{\partial}_t\tilde{e}^n_{\mathrm{f},k}\|^2_{H^1})&\\ +(\frac{1}{2}-\frac{C_{\mathrm{co}}k^{d/2}\xi^k}{2})\|\bar{\partial}_t\eta^n_\theta\|^2+(\kappa\nabla\eta^n_\theta,\nabla\bar{\partial}_t\eta^n_\theta)\leq C\|\bar{\partial}_t\rho^n\|^2,&\end{aligned}$$

where assumption (A4) guarantees that the coefficients are positive. Multiplying by τ, using that $\tau(\kappa\nabla\eta^n_\theta,\nabla\bar{\partial}_t\eta^n_\theta)\geq 1/2(\|\eta^n_\theta\|_\kappa-\|\eta^{n-1}_\theta\|_\kappa)$, and summing over n we derive

$$\begin{aligned}\sum_{j=1}^n\tau(\|\bar{\partial}_t\eta^j_u\|^2_{H^1}+\|\bar{\partial}_t\tilde{e}^j_{\mathrm{f},k}\|^2_{H^1}+\|\bar{\partial}_t\eta^j_\theta\|^2)+\|\eta^n_\theta\|^2_{H^1}&\\ \leq C\sum_{j=1}^n\tau\|\bar{\partial}_t\rho^j\|^2\leq C(H+k^{d/2}\xi^k)\sum_{j=1}^n\tau(\|\bar{\partial}_t f^j\|^2+\|\bar{\partial}_t\theta^j_h\|^2_{H^1}),&\end{aligned}$$

where we have used that $\eta^0_\theta=0$, the bound (2.11), and (5.8)-(5.9). We can now apply Theorem 3.3. Thus, the lemma follows for $\|\theta^n_h-\theta^n_{\mathrm{ms},k}\|_{H^1}$. Moreover, this bounds the last terms in (5.10), which completes the proof. □

LEMMA 5.4. Assume that $f=0$ and $g=0$, and that (A4) holds. Let $\{u^n_h\}_{n=1}^N$ and $\{\theta^n_h\}_{n=1}^N$ be the solutions to the system (3.1)-(3.2) and $\{\tilde{u}^n_{\mathrm{ms},k}\}_{n=1}^N$ and $\{\theta^n_{\mathrm{ms},k}\}_{n=1}^N$ be the solutions to (4.16)-(4.18). For $n\in\{1,...,N\}$ we have

$$\|u^n_h-\tilde{u}^n_{\mathrm{ms},k}\|_{H^1}+t_n^{1/2}\|\theta^n_h-\theta^n_{\mathrm{ms},k}\|_{H^1}\leq C(H+k^{d/2}\xi^k)\|\theta^0_h\|_{H^1}. \tag{5.12}$$

PROOF. As in the proof of Lemma 5.3 we split the error into two parts

$$u^n_h-\tilde{u}^n_{\mathrm{ms},k}=\tilde{\rho}^n_u+\tilde{\eta}^n_u,\quad \theta^n_h-\theta^n_{\mathrm{ms},k}=\rho^n_\theta+\eta^n_\theta,$$

where Lemma 4.2 and Theorem 3.3 gives

$$\begin{aligned}\|\rho^n_\theta\|_{H^1}&\leq C(H+k^{d/2}\xi^k)\|-\bar{\partial}_t\theta^n_h-\nabla\cdot\bar{\partial}_t u^n_h\|\leq C(H+k^{d/2}\xi^k)t_n^{-1/2}\|\theta^0_h\|_{H^1},\\ \|\tilde{\rho}^n_u\|_{H^1}&\leq C(H+k^{d/2}\xi^k)\|\theta^n_h\|_{H^1}\leq C(H+k^{d/2}\xi^k)\|\theta^0_h\|_{H^1}.\end{aligned}$$

Now, note that (5.4) and (5.11) holds also when $f=0$ and $g=0$. In particular, (5.11) holds also for $n=1$ due to the definition of $u^0_{\mathrm{ms},k}$ and u^0_h in (4.19) and

(3.3) respectively. By choosing $v_2 = \bar{\partial}_t \eta_\theta^n$ and adding the resulting equations we derive

$$c_\sigma \|\bar{\partial}_t \eta_u^n\|_{H^1}^2 + \|\bar{\partial}_t \eta_\theta^n\|^2 + (\kappa \nabla \eta_\theta^n, \nabla \bar{\partial}_t \eta_\theta^n) + (\sigma(\bar{\partial}_t \tilde{e}_{\mathrm{f},k}^n) : \epsilon(\bar{\partial}_t \eta_u^n)) \\ + (\alpha \nabla \cdot \bar{\partial}_t \tilde{e}_{\mathrm{f},k}^n, \bar{\partial}_t \eta_\theta^n) \le \|\bar{\partial}_t \rho^n\| \|\bar{\partial}_t \eta_\theta^n\|.$$

Recall $\rho^n = \rho_\theta^n + \alpha \nabla \cdot \tilde{\rho}_u^n$. As in the proof of Lemma 5.3 we get from Lemma 5.2

$$(\alpha \nabla \cdot \bar{\partial}_t \tilde{e}_{\mathrm{f},k}^n, \bar{\partial}_t \eta_\theta^n) \ge -C_{\mathrm{co}} k^{d/2} \xi^k \|\bar{\partial}_t \tilde{e}_{\mathrm{f},k}^n\|_{H^1} \|\bar{\partial}_t \eta_\theta^n\| + (\sigma(\bar{\partial}_t \tilde{e}_{\mathrm{f},k}^n) : \varepsilon(\bar{\partial}_t \tilde{e}_{\mathrm{f},k}^n)).$$

and from (4.15)

$$(\sigma(\tilde{e}_{\mathrm{f},k}^n) : \varepsilon(\bar{\partial}_t \eta_u^n)) \ge -C_{\mathrm{ort}} k^{d/2} \xi^k \|\bar{\partial}_t \tilde{e}_{\mathrm{f},k}^n\|_{H^1} \|\bar{\partial}_t \eta_u^n\|_{H^1}.$$

Hence, we have

$$(c_\sigma - \frac{C_{\mathrm{ort}} k^{d/2} \xi^k}{2}) \|\bar{\partial}_t \eta_u^n\|_{H^1}^2 + (c_\sigma - \frac{(C_{\mathrm{ort}} + C_{\mathrm{co}}) k^{d/2} \xi^k}{2}) \|\bar{\partial}_t \tilde{e}_{\mathrm{f},k}^n\|_{H^1}^2 \\ + (\frac{1}{2} - \frac{C_{\mathrm{co}} k^{d/2} \xi^k}{2}) \|\bar{\partial}_t \eta_\theta^n\|^2 + (\kappa \nabla \eta_\theta^n, \nabla \bar{\partial}_t \eta_\theta^n) \le \|\bar{\partial}_t \rho^n\|^2,$$

and assumption (A4) guarantees that the coefficients are positive. Multiplying by τt_n^2, using that $\tau(\kappa \nabla \eta_\theta^n, \nabla \bar{\partial}_t \eta_\theta^n) \ge 1/2(\|\eta_\theta^n\|_\kappa^2 - \|\eta_\theta^{n-1}\|_\kappa^2)$ and $t_n^2 - t_{n-1}^2 \le 3\tau t_{n-1}$, for $n \ge 2$, now give

$$C\tau t_n^2 (\|\bar{\partial}_t \eta_u^n\|_{H^1}^2 + \|\bar{\partial}_t \tilde{e}_{\mathrm{f},k}^n\|_{H^1}^2 + \|\bar{\partial}_t \eta_\theta^n\|^2) + \frac{t_n^2}{2} \|\eta_\theta^n\|_\kappa^2 - \frac{t_{n-1}^2}{2} \|\eta_\theta^{n-1}\|_\kappa^2 \\ \le C\tau t_n^2 \|\bar{\partial}_t \rho^n\|^2 + C\tau t_{n-1} \|\eta_\theta^{n-1}\|_\kappa^2.$$

Note that this inequality also holds for $n = 1$, since $\eta_\theta^0 = 0$ (recall $\theta_{\mathrm{ms},k}^0 = R_{\mathrm{ms},k}^2 \theta_h^0$). Summing over n gives and using (2.11)

(5.13)
$$C \sum_{j=1}^n \tau t_j^2 (\|\bar{\partial}_t \eta_u^j\|_{H^1}^2 + \|\bar{\partial}_t \tilde{e}_{\mathrm{f},k}^j\|_{H^1}^2 + \|\bar{\partial}_t \eta_\theta^j\|^2) + c_\kappa t_n^2 \|\eta_\theta^n\|_{H^1}^2 \\ \le C \sum_{j=1}^n \tau t_j^2 \|\bar{\partial}_t \rho^j\|^2 + C \sum_{j=1}^{n-1} \tau t_j \|\eta_\theta^j\|_{H^1}^2,$$

and since $f^n = 0$ and $g^n = 0$, Lemma 4.2 and the Aubin-Nitsche trick as in (5.8) together with Theorem 3.3 give

$$\begin{aligned} \|\bar{\partial}_t \rho^j\| &\le \|\bar{\partial}_t \rho_\theta^j\| + \alpha_2 \|\bar{\partial}_t \rho_u^j\|_{H^1} \\ &\le C(H + k^{d/2}\xi^k)(\|\bar{\partial}_t \theta_h^j\|_{H^1} + \|\nabla \cdot \bar{\partial}_t u_h^j\|) \\ &\le C(H + k^{d/2}\xi^k) t_j^{-1} \|\theta_h^0\|_{H^1}. \end{aligned} \tag{5.14}$$

To bound the last sum on the right hand side in (5.13) we choose $v_1 = \bar{\partial}_t \eta_u^n$ and $v_2 = \eta_\theta^n$ in (5.4) and (5.3) and add the resulting equations. This gives

$$
\begin{aligned}
(\sigma(\eta_u^n) : \varepsilon(\bar{\partial}_t \eta_u^n)) &+ (\bar{\partial}_t \eta_\theta^n, \eta_\theta^n) + (\kappa \nabla \eta_\theta^n, \nabla \eta_\theta^n) \\
&= (-\bar{\partial}_t \rho^n, \eta_\theta^n) - (\sigma(\tilde{e}_{\mathrm{f},k}^n) : \varepsilon(\bar{\partial}_t \eta_u^n)) - (\alpha \nabla \cdot \bar{\partial}_t \tilde{e}_{\mathrm{f},k}^n, \eta_\theta^n),
\end{aligned}
$$

where the use of (5.6) and (5.7) gives

$$
\begin{aligned}
(\sigma(\eta_u^n) : \varepsilon(\bar{\partial}_t \eta_u^n)) &+ (\sigma(\tilde{e}_{\mathrm{f},k}^n) : \varepsilon(\bar{\partial}_t \tilde{e}_{\mathrm{f},k}^n)) + (\bar{\partial}_t \eta_\theta^n, \eta_\theta^n) + (\kappa \nabla \eta_\theta^n, \nabla \eta_\theta^n) \\
&\leq \|\bar{\partial}_t \rho^n\| \|\eta_\theta^n\| + C k^{d/2} \xi^k \|\eta_\theta^n\| (\|\bar{\partial}_t \eta_u^n\|_{H^1} + \|\bar{\partial}_t \tilde{e}_{\mathrm{f},k}^n\|_{H^1}).
\end{aligned}
$$

Multiplying by τt_n and using that $t_n - t_{n-1} = \tau$ we get

$$
\begin{aligned}
C\tau t_n \|\eta_\theta^n\|_{H^1}^2 &+ \frac{t_n}{2}(\|\eta_u^n\|_\sigma^2 + \|\tilde{e}_{\mathrm{f},k}^n\|_\sigma^2) - \frac{t_{n-1}}{2}(\|\eta_u^{n-1}\|_\sigma^2 + \|\tilde{e}_{\mathrm{f},k}^{n-1}\|_\sigma^2) \\
&+ \frac{t_n}{2}\|\eta_\theta^n\|^2 - \frac{t_{n-1}}{2}\|\eta_\theta^{n-1}\|^2 \\
&\leq C t_n \tau (\|\bar{\partial}_t \rho^n\| \|\eta_\theta^n\| + k^{d/2} \xi^k \|\eta_\theta^n\| (\|\bar{\partial}_t \eta_u^n\|_{H^1} + \|\bar{\partial}_t \tilde{e}_{\mathrm{f},k}^n\|_{H^1}) \\
&\quad + C\tau (\|\eta_u^{n-1}\|_\sigma^2 + \|\tilde{e}_{\mathrm{f},k}^{n-1}\|_\sigma^2 + \|\eta_\theta^{n-1}\|^2) \\
&\leq C t_n^2 \tau \|\bar{\partial}_t \rho^n\|^2 + C_y t_n^2 \tau k^d \xi^{2k} (\|\bar{\partial}_t \eta_u^n\|_{H^1}^2 + \|\bar{\partial}_t \tilde{e}_{\mathrm{f},k}^n\|_{H^1}^2) \\
&\quad + C\tau (\|\tilde{\eta}_u^{n-1}\|_\sigma^2 + \|\tilde{e}_{\mathrm{f},k}^{n-1}\|_\sigma^2 + \|\eta_\theta^{n-1}\|^2 + \|\eta_\theta^n\|^2),
\end{aligned}
$$

where we have used Young's (weighted) inequality on the form, $\tau t_n ab \leq \tau t_n^2 a^2 + \tau b^2/4$, in the last step. For the second term we have used the inequality with an additional C_y, i.e. $\tau t_n ab \leq C_y \tau t_n^2 a^2 + (4C_y)^{-1} \tau b^2$. Note that C_y can be made arbitrarily small. Summing over n and using (2.10) now gives

$$
\begin{aligned}
C \sum_{j=1}^n \tau t_j \|\eta_\theta^j\|_{H^1}^2 &+ \frac{c_\sigma t_n}{2}(\|\eta_u^n\|_{H^1}^2 + \|\tilde{e}_{\mathrm{f},k}^n\|_{H^1}^2) + \frac{t_n}{2}\|\eta_\theta^n\|^2 \\
&\leq C \sum_{j=1}^n \tau t_j^2 \|\bar{\partial}_t \rho^j\|^2 + C_y k^d \xi^{2k} \sum_{j=1}^n \tau t_j^2 (\|\bar{\partial}_t \eta_u^j\|_{H^1}^2 + \|\bar{\partial}_t \tilde{e}_{\mathrm{f},k}^j\|_{H^1}^2)) \\
&\quad + C \sum_{j=0}^n \tau (\|\eta_u^j\|_{H^1}^2 + \|\tilde{e}_{\mathrm{f},k}^j\|_{H^1}^2 + \|\eta_\theta^j\|^2).
\end{aligned}
\tag{5.15}
$$

We can now use (5.13) to deduce

$$
\sum_{j=1}^n \tau t_j^2 (\|\bar{\partial}_t \eta_u^j\|_{H^1}^2 + \|\bar{\partial}_t \tilde{e}_{\mathrm{f},k}^j\|_{H^1}^2) \leq C \sum_{j=1}^n \tau t_j^2 \|\bar{\partial}_t \rho^j\|^2 + C \sum_{j=1}^{n-1} \tau t_j \|\eta_\theta^j\|_{H^1}^2.
$$

Using this in (5.15) gives

$$C\sum_{j=1}^{n}\tau t_j\|\eta_\theta^j\|_{H^1}^2+\frac{c_\sigma t_n}{2}(\|\eta_u^n\|_{H^1}^2+\|\tilde{e}_{\mathrm{f},k}^n\|_{H^1}^2)+\frac{t_n}{2}\|\eta_\theta^n\|^2 \tag{5.16}$$
$$\leq C\sum_{j=1}^{n}\tau t_j^2\|\bar{\partial}_t\rho^j\|^2+C_yk^d\xi^{2k}\sum_{j=1}^{n}\tau t_j\|\eta_\theta^j\|_{H^1}^2$$
$$+C\sum_{j=0}^{n}\tau(\|\eta_u^j\|_{H^1}^2+\|\tilde{e}_{\mathrm{f},k}^j\|_{H^1}^2+\|\eta_\theta^j\|^2).$$

Since C_y now can be made arbitrarily small the term $C_yk^d\xi^{2k}\sum_{j=1}^n\tau t_j\|\eta_\theta^j\|_{H^1}^2$ can be moved to the left hand side. To estimate the last sum on the right hand side in (5.16) we multiply (5.4) by τ and sum over n to get

$$(\eta_\theta^n-\eta_\theta^0,v_2)+(\kappa\nabla\sum_{j=1}^{n}\tau\eta_\theta^j,\nabla v_2)+(\alpha\nabla\cdot\tilde{\eta}_u^n-\tilde{\eta}_u^0,v_2)=(-\rho^n+\rho^0,v_2), \tag{5.17}$$

where we note that $\eta_\theta^0=0$ and $\tilde{\eta}_u^0=0$. By choosing $v_1=\eta_u^n$ in (5.3) and $v_2=\eta_\theta^n$ in (5.17) and adding the resulting equations we get

$$c_\sigma\|\eta_u^n\|_{H^1}^2+\|\eta_\theta^n\|^2+(\kappa\sum_{j=1}^{n}\tau\nabla\eta_\theta^j,\nabla\eta_\theta^n)$$
$$\leq\|-\rho^n+\rho^0\|\|\eta_\theta^n\|-(\sigma(\tilde{e}_{\mathrm{f},k}^n):\varepsilon(\eta_u^n))-(\alpha\nabla\cdot\tilde{e}_{\mathrm{f},k}^n,\eta_\theta^n).$$
$$\leq\|-\rho^n+\rho^0\|\|\eta_\theta^n\|+C_{\mathrm{ort}}k^{d/2}\xi^k\|\tilde{e}_{\mathrm{f},k}^n\|_{H^1}\|\eta_u^n\|_{H^1}+C_{\mathrm{co}}k^{d/2}\xi^k\|\tilde{e}_{\mathrm{f},k}^n\|_{H^1}\|\eta_\theta^n\|$$
$$-c_\sigma\|\tilde{e}_{\mathrm{f},k}^n\|_{H^1}^2,$$

where we have used the almost orthogonal property (4.15) and Lemma 4.4. We conclude that

$$(c_\sigma-\frac{C_{\mathrm{ort}}k^{d/2}\xi^k}{2})\|\eta_u^n\|_{H^1}^2+(c_\sigma-\frac{(C_{\mathrm{ort}}+C_{\mathrm{co}})k^{d/2}\xi^k}{2})\|\tilde{e}_{\mathrm{f},k}^n\|_{H^1}^2 \tag{5.18}$$
$$+(\frac{1}{2}-\frac{C_{\mathrm{co}}k^{d/2}\xi^k}{2})\|\eta_\theta^n\|^2+(\kappa\sum_{j=1}^{n}\tau\nabla\eta_\theta^j,\nabla\eta_\theta^n)\leq C\|-\rho^n+\rho^0\|^2,$$

and assumption (A4) guarantees positive coefficients. Now, note that we have the bound

$$\Big(\kappa\sum_{j=1}^{n}\tau\nabla\eta_\theta^j,\nabla\eta_\theta^n\Big)=\Big(\kappa\sum_{j=1}^{n}\tau\nabla\eta_\theta^j,\bar{\partial}_t\Big(\sum_{j=1}^{n}\tau\nabla\eta_\theta^j\Big)\Big)$$
$$\geq\frac{1}{2\tau}\Big(\|\sum_{j=1}^{n}\tau\eta_\theta^j\|_\kappa^2-\|\sum_{j=1}^{n-1}\tau\eta_\theta^i\|_\kappa^2\Big),$$

with the convention that $\sum_{j=1}^{0}\tau\eta_\theta^j=0$. Multiplying (5.18) by τ, summing over n, and using (2.11) thus gives

$$\begin{aligned}(5.19)\qquad \sum_{j=1}^{n}\tau(\|\eta_u^j\|_{H^1}^2+\|\tilde{e}_{\mathrm{f},k}^j\|_{H^1}^2+\|\eta_\theta^j\|^2)+\frac{c_\kappa}{2}\|\sum_{j=1}^{n}\tau\eta_\theta^j\|_{H^1}^2\\ \le C\sum_{j=1}^{n}\tau\|-\rho^j+\rho^0\|^2\\ \le C(H+k^{d/2}\xi^k)^2\sum_{j=1}^{n}\tau\|\theta_h^0\|_{H^1}^2\le C(H+k^{d/2}\xi^k)^2t_n\|\theta_h^0\|_{H^1}^2.\end{aligned}$$

Here we have used the Aubin-Nitsche duality argument, Lemma 4.2 and Lemma 3.3 to deduce

$$\begin{aligned}\|\rho^j\|\le\|\rho_\theta^j\|+C\|\rho_u^j\|_{H^1}&\le C(H+k^{d/2}\xi^k)(\|\rho_\theta^j\|_{H^1}+\|\theta_h^n\|_{H^1})\\ &\le C(H+k^{d/2}\xi^k)\|\theta_h^n\|_{H^1}\le C(H+k^{d/2}\xi^k)\|\theta_h^0\|_{H^1},\quad j\ge0.\end{aligned}$$

Combining (5.13), (5.14), (5.16), and (5.19) we get

$$t_n^2\|\eta_\theta^n\|_{H^1}^2+t_n\|\eta_u^n\|_{H^1}^2+t_n\|\tilde{e}_{\mathrm{f},k}^n\|_{H^1}^2\le C(H+k^{d/2}\xi^k)^2t_n\|\theta_h^0\|_{H^1}^2,$$

which completes the proof. □

Proof of Theorem 5.2. Since the problem is linear we can split the solution

$$u_h^n=\bar{u}_h^n+\hat{u}_h^n,\quad \theta_h^n=\bar{\theta}_h^n+\hat{\theta}_h^n,$$

where $\bar{u}_h^n$ and $\bar{\theta}_h^n$ solves (3.1)-(3.2) with $f=0$ and $g=0$ and $\hat{u}_h^n$ and $\hat{\theta}_h^n$ solves (3.1)-(3.2) with $\theta^0=0$. The theorem now follows by applying Lemma 5.3 and Lemma 5.4. □

6. Numerical examples

In this section we perform two numerical examples. For a discussion on how to implement the type of generalized finite element efficiently described in this paper we refer to [9].

The first numerical example models a composite material which is preheated to a fix temperature and at time $t_0=0$ the piece is subject to a cool-down.

The domain is set to be the unit square $\Omega=[0,1]\times[0,1]$ and we assume that the temperature has a homogeneous Dirichlet boundary condition, that is $\Gamma_D^\theta=\partial\Omega$ and $\Gamma_N^\theta=\emptyset$. For the displacement we assume the bottom boundary to be fix and for the remaining part of the boundary we prescribe a homogeneous Neumann condition, that is $\Gamma_D^u=[0,1]\times 0$ and $\Gamma_N^u=\partial\Omega\setminus\Gamma_D^u$.

The composite is assumed to be built up according to Figure 7. The white part in the figure denotes a background material and the black parts an insulated material. The black squares are of size $2^{-5}\times2^{-5}$. We assume that the Lamé coefficients μ and λ take the values μ_1 and λ_1 on the insulated material, and μ_2 and λ_2 on the background material. In this experiment we have set $\mu_1/\mu_2=10$

and $\lambda_1/\lambda_2 = 50$. Similarly, using subscript 1 for the insulated material and subscript 2 for the background material, we set $\alpha_1/\alpha_2 = 10$ and $\kappa = \kappa_i \cdot I$, for $i = 1, 2$, where I is the 2-dimensional identity matrix and $\kappa_1/\kappa_2 = 10$. Furthermore, we have chosen to set $f = [0, 0]^\intercal$ (no external body forces) and $g = -10$.

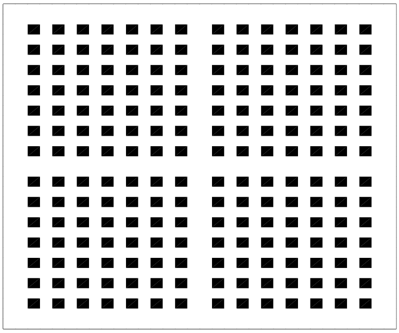


Figure 7. Composite material on the unit square. One black square is of size $2^{-5} \times 2^{-5}$.

The initial data must be zero on the boundary Γ_D^θ, so we have chosen to put $\theta^0 = 500x(1-x)y(1-y)$ and θ_h^0 to the L_2-projection of θ^0 to V_h^2. For the generalized finite element solution we have chosen $\theta_{\mathrm{ms},k}^0 = R_{\mathrm{ms},k}^2 \theta_h^0$ and $\tilde{u}_{\mathrm{ms},k}^0$ is given by (4.19).

The domain is discretized using a uniform triangulation. The reference solution is computed on a mesh of $h = \sqrt{2} \cdot 2^{-6}$ which resolves the fine parts (the black squares) in the material. The generalized finite element method (GFEM) in (4.16)-(4.18) is computed for five decreasing values of the mesh size, namely, $H = \sqrt{2} \cdot 2^{-1}, \sqrt{2} \cdot 2^{-2}, ..., \sqrt{2} \cdot 2^{-5}$, with the patch sizes $k = 1, 1, 2, 2, 3$. For comparison, we also compute the corresponding classical finite element (FEM) solution on the coarse meshes using continuous piecewise affine polynomials for both spaces (P1-P1). The solutions satisfies (3.1)-(3.2) with h replaced by H and are denoted u_H^n and θ_H^n respectively for $n = 1, ..., N$. When computing these solutions we have evaluated the integrals exactly to avoid quadrature errors.

We have chosen to set $T = 1$ and $\tau = 0.05$ for all values of H and for the reference solution. The solutions are compared at the time point N.

Note that the implementation of the corrections $u_{\mathrm{f},k}^{n,K}$ in (4.18) given by

$$(\sigma(u_{\mathrm{f},k}^{n,K}) : \varepsilon(w_1)) - (\alpha \theta_{\mathrm{ms},k}^n, \nabla \cdot w_1)_K = 0, \quad \forall w_1 \in V_{\mathrm{f}}^1(\omega_k(K)),$$

should *not* be computed explicitly at each time step. It is more efficient to compute x_y^K, given by

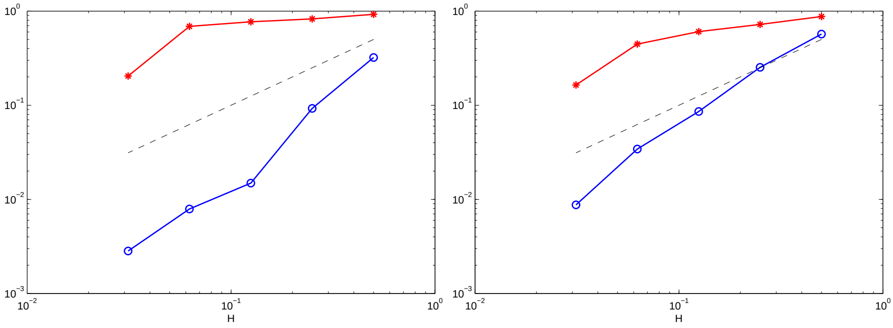
$$(\sigma(x_y^K) : \varepsilon(w_1)) - (\alpha(\lambda_y^2 - R_{\mathrm{f},k}^2 \lambda_y^2), \nabla \cdot w_1)_K = 0, \quad \forall w_1 \in V_{\mathrm{f}}^1(\omega_k(K)),$$

where $\{(\cdot,y) \in \mathcal{N} : \lambda_y^2 - R_{\mathrm{f},k}^2 \lambda_y^2\}$ is the basis for $V_{\mathrm{ms},k}^2$. Now, since $\theta_{\mathrm{ms},k}^n = \sum_y \beta_y^n(\lambda_y^2 - R_{\mathrm{f},k}^2\lambda_y^2)$, we have the identity

$$u_{\mathrm{f},k}^n = \sum_K u_{\mathrm{f},k}^{n,K} = \sum_K \sum_y \beta_y^n x_y^K.$$

With this approach, we only need to compute x_y^K once before solving for the system (4.16)-(4.17) for $n = 1, ..., N$.

The relative errors in the H^1-seminorm $\|\nabla\cdot\|$ are shown in Figure 8. The left graph shows the relative errors for the displacement, $\|\nabla(\tilde{u}_{\mathrm{ms},k}^N - u_h^N)\|/\|\nabla u_h^N\|$ and $\|\nabla(u_H^N - u_h^N)\|/\|\nabla u_h^N\|$. The right graph shows the error for the temperature $\|\nabla(\theta_{\mathrm{ms},k}^N - \theta_h^N)\|/\|\nabla\theta_h^N\|$ and $\|\nabla(\theta_H^N - \theta_h^N)\|/\|\nabla\theta_h^N\|$. As expected the generalized finite element shows convergence of optimal order and outperforms the classical finite element.



(A) Displacement u

(B) Temperature θ

FIGURE 8. Relative errors using GFEM (blue ◦) and P1-P1 FEM (red $*$) for the linear thermoelasticity problem plotted against the mesh size H. The dashed line is H.

The second example shows the importance of the additional correction (4.18), which is designed to handle multiscale behavior in the coefficient α. The computational domain, the spatial and the time discretization, and the patch sizes remain the same as in the first example. However, we let $\Gamma_D = \partial\Omega$ and $\Gamma_N = \emptyset$ in this case.

To test the influence of α we let the other coefficients be constants, $\mu = \lambda = 1$ and $\kappa = I$, where the I is the 2-dimensional identity matrix. The coefficient α takes values between 0.1 and 10 according to Figure 9. The boxes are of size $2^{-5} \times 2^{-5}$ and, hence, the reference mesh of size $h = \sqrt{2} \cdot 2^{-6}$ is sufficiently small to resolve the variations in α.

The initial data is set to $\theta^0 = x(1-x)y(1-y)$ and θ_h^0 is the L_2-projection of θ^0 onto V_h^2. For the generalized finite element solution we have chosen $\theta_{\mathrm{ms},k}^0 = R_{\mathrm{ms},k}^2\theta_h^0$ and $\tilde{u}_{\mathrm{ms},k}^0$ is given by (4.19), as in our first example. Furthermore, we have chosen to set $f = [1\ 1]^\intercal$ and $g = 10$.

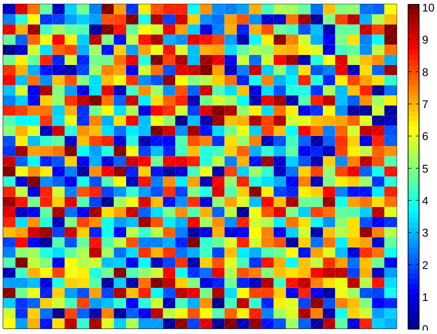


Figure 9. A plot of the coefficient α.

The generalized finite element method (GFEM) in (4.16)-(4.18) is computed for the five decreasing values of the mesh size used in the first example. For comparison, we compute the generalized finite element without the additional correction on $u^n_{\mathrm{ms},k}$. In this case the system (4.16)-(4.18) simplifies to

$$\begin{aligned}(\sigma(u^n_{\mathrm{ms},k}):\varepsilon(v_1))-(\alpha\theta^n_{\mathrm{ms},k},\nabla\cdot v_1)&=(f^n,v_1),\quad \forall v_1\in V^1_{\mathrm{ms},k},\\ (\bar{\partial}_t\theta^n_{\mathrm{ms},k},v_2)+(\kappa\nabla\theta^n_{\mathrm{ms},k},\nabla v_2)+(\alpha\nabla\cdot\bar{\partial}_t u^n_{\mathrm{ms},k},v_2)&=(g^n,v_2).\quad \forall v_2\in V^2_{\mathrm{ms},k}\end{aligned}$$

The relative errors in the H^1-seminorm are shown in Figure 8. The graph shows the errors for the displacement with correction for α, $\|\nabla(\tilde{u}^N_{\mathrm{ms},k}-u^N_h)\|/\|\nabla u^N_h\|$ and the error without correction for α $\|\nabla(u^N_{\mathrm{ms},k}-u^N_h)\|/\|\nabla u^N_h\|$. As expected the GFEM with correction for α shows convergence of optimal order and outperforms the GFEM without correction for α. This is due to the fact that the constant in (4.13) (and hence also the constant in Theorem 5.2) depends on the variations in α.

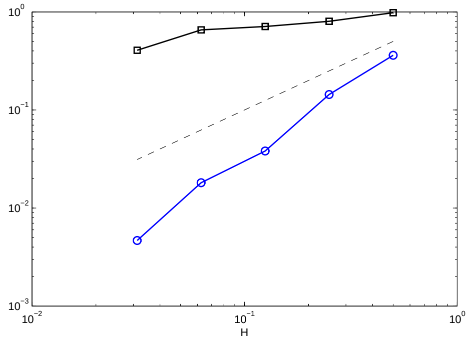


Figure 10. Relative errors for the displacement u using GFEM with correction for α (blue $\circ$) and GFEM without correction for α (black $\square$) for the linear thermoelasticity problem plotted against the mesh size H. The dashed line is H.

References

[1] A. Abdulle and P. Henning: *Localized orthogonal decomposition method for the wave equation with a continuum of scales*, to appear in Math. Comp., 2016+.

[2] K.T. Andrews and P. Shi and M. Shillor and S. Wright: *Thermoelastic contact with Barber's heat exchange condition*, Appl. Math. Optim. 28 (1993), no. 1, p. 11–48.

[3] I. Babuška and R. Lipton: *Optimal local approximation spaces for generalized finite element methods with application to multiscale problems*, Multiscale Model. Simul. 9 (2011), no. 1, p. 373–406.

[4] I. Babuška and J. E. Osborn: *Generalized finite element methods: their performance and their relation to mixed methods*, SIAM J. Numer. Anal. 20 (1983), no. 3, 1983.

[5] M. A. Biot: *General theory of three-dimensional consolidation*, J. Appl. Phys., 18 (1941), no. 2, p. 155–164.

[6] M. A. Biot: *Thermoelasticity and irreversible thermodynamics*, J. Appl. Phys., 27 (1956), p. 240–253.

[7] P. G. Ciarlet: *Mathematical elasticity. Vol. I*, vol. 20 of *Studies in Mathematics and its Applications*, North-Holland Publishing Co., Amsterdam, 1988.

[8] C. M. Dafermos: *On the existence and the asymptotic stability of solutions to the equations of linear thermoelasticity*, Arch. Rational Mech. Anal. 29 (1968), p. 241–271.

[9] Ch. Engwer and P. Henning and A. Målqvist and D. Peterseim: *Efficient implementation of the Localized Orthogonal Decomposition method*, Submitted.

[10] A. Ern and S. Meunier: *A posteriori error analysis of Euler-Galerkin approximations to coupled elliptic-parabolic problems*, M2AN Math. Model. Numer. Anal. 43 (2009), no. 2, p. 353 – 375.

[11] P. Henning and A. Målqvist: *Localized orthogonal decomposition techniques for boundary value problems*, SIAM J. Sci. Comput. 36 (2014), no. 4, p. A1609–A1634.

[12] P. Henning, A. Målqvist and D. Peterseim: *A localized orthogonal decomposition method for semi-linear elliptic problems*, ESAIM Math. Model. Numer. Anal. 48 (2014), no. 5, p. 1331–1349.

[13] P. Henning and A. Persson: *A multiscale method for linear elasticity*

reducing Poisson locking, Submitted.

[14] T. J. R. Hughes, G. R. Feijóo, L. Mazzei, and J-B. Quincy: *The variational multiscale method—a paradigm for computational mechanics*, Comput. Methods Appl. Mech. Engrg. 166 (1998), no. 1-2, p. 3–24.

[15] M. G. Larson and A. Målqvist: *Adaptive variational multiscale methods based on a posteriori error estimation: energy norm estimates for elliptic problems*, Comput. Methods Appl. Mech. Engrg. 196 (2007), no. 21-24, p. 2313–2324.

[16] A. Målqvist and A. Persson: *Multiscale techniques for parabolic equations*, Submitted, 2015.

[17] A. Målqvist and D. Peterseim: *Computation of eigenvalues by numerical upscaling*, Numer. Math. 130 (2015), no. 2, p. 337–361.

[18] A. Målqvist and D. Peterseim: *Localization of elliptic multiscale problems*, Math. Comp. 83 (2014), no. 290, p. 2583–2603.

[19] D. Peterseim: *Variational Multiscale Stabilization and the Exponential Decay of Fine-scale Correctors*, Submitted.

[20] P. J. Phillips and M. F. Wheeler: *A coupling of mixed and discontinuous Galerkin finite-element methods for poroelasticity*, Computational Geosciences 12 (2008), no. 4, p. 417–435.

[21] P. Shi and M. Shillor: *Existence of a solution to the N-dimensional problem of thermoelastic contact*, Comm. Partial Differential Equations 17 (1992), no. 9-10, p. 1597–1618.

[22] R. E. Showalter: *Diffusion in poro-elastic media*, J. Math. Anal. Appl. 251 (2000), no. 1, p.310–340.

[23] X. Xu: *The N-dimensional quasistatic problem of thermoelastic contact with Barber's heat exchange conditions*, Adv. Math. Sci. Appl. 6 (1996), no. 2, p. 559–587.

[24] A. Ženíšek: *Finite element methods for coupled thermoelasticity and coupled consolidation of clay*, RAIRO Anal. Numér. 18 (1984), no. 2, p. 183–205.

[25] A. Ženíšek: *The existence and uniqueness theorem in Biot's consolidation theory*, Apl. Mat. 29 (1984), no. 3, p. 194–211.