Numerical homogenization for nonlinear strongly monotone problems

Barbara Verfürth

KARLSRUHE INSTITUTE OF TECHNOLOGY

CRC 1173

Wa∨e
phenomena

Participating universities

Universität Stuttgart

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

TU
WIEN

Funded by

DFG

Numerical homogenization for nonlinear strongly monotone problems[*]

Barbara Verfürth[†]

Abstract. In this work we introduce and analyze a new multiscale method for strongly nonlinear monotone equations in the spirit of the Localized Orthogonal Decomposition. A problem-adapted multiscale space is constructed by solving linear local fine-scale problems which is then used in a generalized finite element method. The linearity of the fine-scale problems allows their localization and, moreover, makes the method very efficient to use. The new method gives optimal a priori error estimates up to linearization errors beyond periodicity and scale separation and without assuming higher regularity of the solution. The effect of different linearization strategies is discussed in theory and practice. Several numerical examples including stationary Richards equation confirm the theory and underline the applicability of the method.

Key words. multiscale method; numerical homogenization; nonlinear monotone problem; a priori error estimates

AMS subject classifications. 65N15, 65N30, 35J60, 74Q15

1. Introduction

Linear constitutive laws like Hooke's law in mechanics, Ohm's law in electromagnetics, or Darcy's law in fluid flow are very popular, but they are often not accurate enough in practical applications, for instance for high intensities. Instead, nonlinear effects in the constitutive laws have to be taken into account which are often experimentally found and determined, see [40] for a general overview. In this article, we consider as model problem the following nonlinear monotone elliptic equation

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

where the exact assumptions as well as boundary conditions are specified later. It is a representative model problem for quasilinear partial differential equations (PDEs) as they occur in mean curvature flow or for non-Newtonian fluids. Note that this class of problems is far more challenging than semilinear PDEs, where the nonlinearities do not appear in the highest derivative. The transition from linear to nonlinear problems comes with huge additional challenges for the numerical treatment and analysis. As an illustrating example we mention optimal order L^2-estimates for the finite element method: The classical Aubin-Nitsche trick for linear problems is not applicable, so that, for a long time, only optimal order estimates in the energy norm [10] were known, see [5] and the discussion therein. A similar observation applies to the effect of numerical integration, see [17].

[†]Institut für Angewandte und Numerische Mathematik, Karlsruher Institut für Technologie, Englerstr. 2, D-76131 Karlsruhe

With the view on practical applications such as fluid flow or elasticity, we do not only have to consider nonlinear constitutive laws as discussed above, but also have to consider (spatial) multi-scale features in the material coefficients (here, in A). For instance, a fluid such as groundwater flows over large distances, while the properties of the soil changes over small distances and thereby influences the overall groundwater flow, see, e.g., [37]. Hence, for applications such as the (quasi-linear) porous medium equation, A is subject to rapid variations and/or discontinuities on fine spatial scales or even a cascade of (non-separable) scales. This coincidence of multiscale features and nonlinear material laws makes the problem intractable for standard (problem-independent) methods. For example, the standard finite element method [2, 10, 17] will only give optimal convergence in the asymptotic regime, i.e., if the mesh resolves all features and scales present in the coefficient, which is prohibitively expensive even with today's computational resources.

In the case of spatially periodic A (with period $\varepsilon \ll 1$), homogenization results using two-scale convergence [7, 31] prove that the solutions of the above model problem converge to the solution of an again monotone elliptic (homogenized) problem for $\varepsilon \to 0$. The nonlinear effective diffusion tensor can be computed by solving nonlinear so-called cell problems. The (finite element) heterogeneous multiscale method is inspired by this analytical process and it is studied successfully for nonlinear problems in a series of papers [1, 3, 4, 6, 21, 25]. In most cases, the macroscopic nonlinear form involves nonlinear reconstruction operators which require the solution of nonlinear cell problems at each macroscopic quadrature point and thereby incorporate the necessary fine-scale information. For parabolic equations, [4] linearizes the macroscopic and cell computations using information from the previous time step. The sparse multiscale FEM [27] is also motivated by homogenization results and tries to reduce the complexity of solving cell problems and a homogenized equation by the introduction of sparse approximations. Another idea to cope with multiscale problems is to modify or enrich the standard finite element basis by problem-adapted functions. This idea is used for instance in the (generalized) multiscale finite element method, for which nonlinear problems are discussed in [9, 12, 13]. Again nonlinear problems have to be solved locally to construct the problem-adapted functions.

The main contribution of this article is the introduction of a new multiscale method for nonlinear monotone elliptic problems and its numerical analysis. The idea is to construct a multiscale space by solving local fine-scale problems in the spirit of the Localized Orthogonal Decomposition (LOD) [32, 35]. In contrast to the above discussed methods, the basis construction only requires the solution of *linear* problems and hence is embarrassingly easy to implement and use. Moreover, this linearization idea drastically reduces the computational effort for generating a problem-dependent (multiscale) basis and thereby provides a conceptually new view on the treatment of nonlinear multiscale problems. We derive optimal convergence rates (with respect to the mesh size H) up to linearization errors without any assumption on the regularity of the exact solution or special properties such as periodicity or scale separation for the coefficient. The occurring linearization errors and resulting possible choices of the linearization are discussed and compared. Extensive numerical experiments show the good performance of the method in agreement with the theoretical estimates. We study periodic as well as completely random multiscale coefficients and also include a model for stationary Richards equation with a high contrast channel. Because of the relation between the LOD and homogenization theory for periodic coefficients on suitable meshes [18], our numerical experiments with locally periodic diffusion tensors indicate that the presented idea of linearizing the fine-scale problems might be transferable to the Heterogeneous Multiscale Method in [21, 28]. Besides several linear problem classes, the LOD has already been studied for semilinear equations [23] and a nonlinear eigenvalue problem related to the Gross-Pitaevskii equation [24]. These problems, however, are only semilinear and can therefore be handled easier. Yet, we emphasize that these previous works can be re-interpreted in the current framework. We mention the close connections of the LOD to (analytical) homogenization [18, 36], domain decomposition iterative solvers [29, 30, 36], and

so-called gamblets [33, 34]. Hence, the current approach can give interesting and useful insights in these areas for nonlinear problems as well, a direction that is not pursued further in this work.

The article is organized as follows: Section 2 introduces the setting and the standard finite element discretization. We introduce the multiscale method including linearization and localization in Section 3. The arising errors are analyzed in Section 4. Finally, we present several numerical experiments confirming our theory and showing possible applications in Section 5.

2. Problem formulation and discretization

In this section we formulate the considered model problem and introduce necessary finite element prerequisites. We use standard notation on Sobolev spaces. Throughout the whole article, let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain. For a subdomain $D \subset \Omega$, let $\| \cdot \|_{0,D}$, $\| \cdot \|_{1,D}$, and $| \cdot |_{1,D}$ denote the standard $L^2(D)$-norm, $H^1(D)$-norm, and $H^1(D)$-semi norm, respectively. Furthermore, $(\cdot, \cdot)_D$ denotes the standard L^2 scalar product on D. We will omit the subscript D if it equals the full computational domain Ω.

2.1. Model problem

We consider the following nonlinear elliptic problem: Find $u : \Omega \to \mathbb{R}$ such that

$$
\begin{aligned}
-\nabla \cdot \big(A(x, \nabla u)\big) &= f \quad \text{in } \Omega, \\
u &= 0 \quad \text{on } \partial\Omega
\end{aligned}
\tag{2.1}
$$

with a right-hand side $f \in L^2(\Omega)$. The corresponding weak formulation, with which we will work in the following, reads: Find $u \in H_0^1(\Omega)$ such that

$$
\mathcal{B}(u; v) := \big(A(x, \nabla u), \nabla v\big)_\Omega = (f, v)_\Omega \qquad \text{for all} \quad v \in H_0^1(D).
\tag{2.2}
$$

For simplicity, we restrict ourselves to homogeneous Dirichlet boundary conditions, but non-homogeneous and Neumann boundary conditions could be treated as well, see [22]. Moreover, we focus on nonlinearities in the highest derivative only, additional (nonlinear) low-order terms can easily be handled as well, cf. [23]. We now specify the assumptions on A, which guarantee existence and uniqueness of a solution.

Assumption 2.1. *The nonlinearity $A : \Omega \times \mathbb{R}^d \to \mathbb{R}^d$ satisfies*

1. *$A(\cdot, \xi) \in L^\infty(\Omega; \mathbb{R}^d)$ for all $\xi \in \mathbb{R}^d$ and $A(x, \cdot) \in C^1(\mathbb{R}^d; \mathbb{R}^d)$ for almost every $x \in \Omega$;*

2. *there is $\Lambda > 0$ such that $|A(x, \xi_1) - A(x, \xi_2)| \leq \Lambda |\xi_1 - \xi_2|$ for almost every $x \in \Omega$ and all $\xi_1, \xi_2 \in \mathbb{R}^d$;*

3. *there is $\lambda > 0$ such that $\big(A(x, \xi_1) - A(x, \xi_2)\big) \cdot (\xi_1 - \xi_2) \geq \lambda |\xi_1 - \xi_2|^2$ for almost every $x \in \Omega$ and all $\xi_1, \xi_2 \in \mathbb{R}^d$;*

4. *$|A(x, 0)| \leq C_0$ for almost every $x \in \Omega$.*

Assumption 2.1 implies that

$$
|\mathcal{B}(v_1; \psi) - \mathcal{B}(v_2; \psi)| \leq \Lambda |v_1 - v_2|_1 \, |\psi|_1 \qquad \text{and} \qquad \mathcal{B}(v; v - \psi) - \mathcal{B}(\psi; v - \psi) \geq \lambda |v - \psi|_1^2
$$

for all $v, v_1, v_2, \psi \in H_0^1(\Omega)$. Therefore, the model problem (2.2) has a unique solution $u \in H_0^1(\Omega)$, which satisfies

$$
|u|_1 \leq \Lambda/\lambda \|f\|_0,
\tag{2.3}
$$

3

see [10, Chapter 5]. As discussed in the introduction, we implicitly assume that A is subject to rapid oscillations or discontinuities on a rather fine scale with respect to the spatial variable x.

We write $a \lesssim b$ in short for $a \leq Cb$ with a constant C independent of the mesh size H and the oversampling parameter m introduced later. However, C may depend on the monotonicity and Lipschitz constants λ, Λ of A (cf. Assumption 2.1).

2.2. Finite element discretizations

We cover Ω with a regular mesh $\mathcal{T}_H$ consisting of simplices; however, a mesh with quadrilaterals would equally be possible. The mesh is assumed to be shape regular in the sense that the aspect ratio of the elements of $\mathcal{T}_H$ is bounded uniformly from below. We introduce the mesh size $H = \max_{T \in \mathcal{T}_H} \operatorname{diam} T$ and assume that this is rather coarse, in particular, $\mathcal{T}_H$ does not resolve the possible heterogeneities in A. We discretize the space $H_0^1(\Omega)$ with the lowest order Lagrange elements over $\mathcal{T}_H$, and denote this space by V_H. This means that $V_H = H_0^1(\Omega) \cap \mathcal{S}^1(\mathcal{T}_H)$, where $\mathcal{S}^1(\mathcal{T}_H)$ denotes the space of element-wise polynomials of total degree ≤ 1.

The standard finite element method now seeks a (discrete) solution $u_H \in V_H$ such that

$$\mathcal{B}(u_H; v_H) = (f, v_H)_\Omega \qquad \text{for all} \quad v_H \in V_H.$$

This results in a nonlinear system which can be (approximatively) solved via an iteration such as Newton's method. If we neglect numerical errors introduced by the inexact solving due to the iteration, it is well-known that the properties of A and Galerkin orthogonality imply

$$|u - u_H|_1 \lesssim \inf_{v_H \in V_H} |u - v_H|_1, \tag{2.4}$$

see [10, Chapter 5]. This quasi-optimality by the way holds for any conforming subset $\tilde{V}_H \subset H_0^1(\Omega)$. For the standard finite element method it is furthermore well-known to have the following (a priori) error estimates with $k > 0$

$$\|u - u_H\|_1 \leq CH^k \|u\|_{H^{1+k}(\Omega)} \qquad \text{and} \qquad \|u - u_H\|_0 \leq CH^{k+1} \|u\|_{H^{1+k}(\Omega)},$$

see [2]. In the above error estimates higher regularity (more than $H_0^1(\Omega)$) of the exact solution is required. However, the regularity of u may be very low for nonlinearities A with spatial discontinuities. Even if the exact solution u satisfies sufficient higher regularity, the corresponding norms $\|u\|_{H^{k+1}(\Omega)}$ depend on spatial derivatives of A which behave like ε^{-q} for some $q \geq 1$ for coefficients varying on a scale ε. In practice this implies that H needs to be at least ε in order to observe the linear convergence in the $H^1(\Omega)$-norm. In other words, for small ε, there is a large pre-asymptotic region where no error estimates are available and the error stagnates (at a high level) in practice.

The goal of the multiscale method presented in Section 3 is to circumvent both issues (higher regularity of the solution and dependence on the variations of A). At the heart of the method is the choice of a suitable interpolation operator and we now introduce the required properties as well as an appropriate example. Let $I_H : H_0^1(\Omega) \to V_H$ denote a bounded local linear projection operator, i.e., $I_H \circ I_H = I_H$, with the following stability and approximation properties for all $v \in H_0^1(\Omega)$

$$|I_H v|_{1,T} \lesssim |v|_{1,\mathrm{N}(T)}, \tag{2.5}$$

$$\|v - I_H v\|_{0,T} \lesssim H|v|_{1,\mathrm{N}(T)}. \tag{2.6}$$

where the constants are independent of H and $\mathrm{N}(T) := \{K \in \mathcal{T}_H : K \cap T \neq \emptyset\}$ denotes the neighborhood of an element. A possible choice (which we use in our implementation of the

method) is to define $I_H := E_H \circ \Pi_H$. On each element $T \in \mathcal{T}_H$, the affine function $(\Pi_H v)|_T$ is the best approximation of $v \in H_0^1(\Omega)$ in the $\|\cdot\|_{0,T}$-norm, and E_H is the averaging operator that maps discontinuous functions in $\mathcal{S}^1(\mathcal{T}_H)$ to V_H by assigning to each free vertex the arithmetic mean of the corresponding function values of the neighboring cells, that is, for any $v \in \mathcal{S}^1(\mathcal{T}_H)$ and any vertex z of $\mathcal{T}_H$,

$$(E_H(v))(z) = \sum_{T \in \mathcal{T}_H, \, z \in T} v|_T(z) \Big/ \operatorname{card}\{K \in \mathcal{T}_H \, , \, z \in K\}.$$

For further details on suitable interpolation operators we refer to [16].

3. Computational multiscale method

In the following, we assume that an interpolation operator $I_H : H_0^1(\Omega) \to V_H$ satisfying the projection property as well as (2.5) and (2.6) is at hand. Abbreviating $W := \ker I_H$, we have the splitting $H_0^1(\Omega) = V_H \oplus W$. The main idea of the Localized Orthogonal Decomposition [32, 35] is to make this splitting problem-dependent. For instance, in the linear elliptic case the splitting is orthogonalized with respect to the energy scalar product. Below, we discuss how this idea can be transferred to the nonlinear case, which turns out to be highly non-trivial. We introduce a linearization procedure in the next subsection which makes the computation of a multiscale space in the spirit of the LOD possible. Afterwards, we present the localized computation of the new multiscale basis functions.

3.1. An ideal method and its linearization

Motivated by linear elliptic equations, one could (naively) try to introduce a Galerkin method over a subset $V_H^{\mathrm{nl,ms}} \subset H_0^1(\Omega)$, i.e., we seek $u_H^{\mathrm{nl,ms}}$ such that

$$\mathcal{B}(u_H^{\mathrm{nl,ms}}; v) = (f, v)_\Omega \qquad \text{for all} \quad v \in V_H^{\mathrm{nl,ms}},$$

where the set $V_H^{\mathrm{nl,ms}}$ is defined via

$$\mathcal{B}(v_H^{\mathrm{nl,ms}}; w) = 0 \qquad \text{for all} \quad v_H^{\mathrm{nl,ms}} \in V_H^{\mathrm{nl,ms}} \quad \text{and all} \quad w \in W. \qquad (3.1)$$

This is the orthogonalization idea behind the original method, see [32, 35]. Due to the quasi-optimality (2.4) and the properties (2.5) and (2.6) of I_H, one obtains the a priori error estimate

$$|u - u_H^{\mathrm{nl,ms}}|_1 \lesssim H\|f\|_0$$

with optimal rate in the mesh size, independent of the regularity of the continuous solution u. This estimate is derived similar to the linear case and we refer to [32, 35] for details. Because of the nonlinearity of $\mathcal{B}$ in its first argument, however, $V_H^{\mathrm{nl,ms}}$ is no longer a linear subspace. To be more precise, we consider the usual construction of $V_H^{\mathrm{nl,ms}}$: It holds $V_H^{\mathrm{nl,ms}} = (\mathrm{id} - \mathcal{Q}^{\mathrm{nl}})V_H$, where $\mathcal{Q}^{\mathrm{nl}} : V_H \to W$ solves

$$\mathcal{B}(v_H - \mathcal{Q}^{\mathrm{nl}} v_H; w) = 0 \qquad \text{for all} \quad w \in W. \qquad (3.2)$$

Here, we clearly see that $\mathcal{Q}^{\mathrm{nl}}$ is a nonlinear operator. Therefore, it is by no means clear whether the proposed multiscale method is at all well defined. Even if this is the case, the method is very complicated as it involves two coupled nonlinear problems, where the one for the generation of

the correction is additionally posed on the fine scale. Hence, we need to simplify in particular the generation of the problem-adapted multiscale basis.

Here, we propose the following simple yet effective linearization approach. We approximate the nonlinearity $A(x, \xi)$ by a function $A_L(x, \xi^*, \xi)$. Here, $A_L : \Omega \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ is affine in its last argument and we call $\xi^* \in \mathbb{R}^d$ the linearization point. We make the following assumption on A_L.

Assumption 3.1. Write $A_L(x, \xi^*, \xi) = \mathbf{A}_L(x, \xi^*)\xi + b_L(x, \xi^*)$ with $\mathbf{A}_L(x, \xi^*) \in \mathbb{R}^{d \times d}$ and $b_L(x, \xi^*) \in \mathbb{R}^d$. We assume that

- $\mathbf{A}_L(\cdot, \xi^*) \in L^\infty(\Omega; \mathbb{R}^{d \times d})$ and $b_L(\cdot, \xi^*) \in L^\infty(\Omega; \mathbb{R}^d)$ for all $\xi^* \in \mathbb{R}^d$;

- $\mathbf{A}_L(x, \xi^*)$ is symmetric for all $\xi^* \in \mathbb{R}^d$;

- there exists $0 < \lambda_L \leq \Lambda_L$ such that
$$\lambda_L |\xi|^2 \leq \mathbf{A}_L(x, \xi^*)\xi \cdot \xi \leq \Lambda_L |\xi|^2 \qquad \text{for all} \quad x \in \Omega, \xi^* \in \mathbb{R}^d, \xi \in \mathbb{R}^d;$$

- $|b_L(x, \xi^*)| \leq C_b$ for all $x \in \Omega$ and all $\xi^* \in \mathbb{R}^d$.

The assumption of symmetry is only made for convenience and to avoid cluttering of notation in the following. Although this linearization model may seem rather abstract, it is commonly used when solving (discrete) nonlinear problems by an iterative method. More precisely, our model for A_L includes Newton-type as well as Kačanov-type linearizations as illustrated in the example, cf. [14].

Example 3.2. Newton-type linearizations are based on a Taylor expansion up to the first order of the nonlinearity around the linearization point. In particular, we approximate $A(x, \xi) \approx A(x, \xi^*) + D_\xi A(x, \xi^*)\nabla(\xi - \xi^*)$, where $D_\xi A$ denotes the Jacobian of A with respect to the second argument. In the notation of Assumption 3.1, this means that $b_L(x, \xi^*) = A(x, \xi^*) - D_\xi A(x, \xi^*)\xi^*$ and $\mathbf{A}_L(x, \xi^*) = D_\xi A(x, \xi^*)$. The assumptions of strict monotonicity and Lipschitz continuity on A (cf. Assumption 2.1) imply that $\mathbf{A}_L$ indeed satisfies Assumption 3.1, see [28, Lemma 6.5.2].

In the case that A takes the form $A(x, \nabla u) = \alpha(x, |\nabla u|^2)\nabla u$, Kačanov-type linearizations are very popular, which "freeze the nonlinearity". In the language of Assumption 3.1, one sets $A_L(x, \xi^*, \xi) := \alpha(x, |\xi^*|^2)\xi$, and, hence, $\mathbf{A}_L(x, \xi^*) = \alpha(x, |\xi^*|^2)$ and $b_L(x, \xi^*) = 0$.

Let us come back to the linearization of our corrector computation. We pick and fix a function $u^* \in H_0^1(\Omega)$ and set $\mathfrak{A} := \mathbf{A}_L(\cdot, \nabla u^*)$. Roughly speaking, this means that we select ∇u^* as our linearization point. We define the *linear* correction operator $\mathcal{Q} : V_H \to W$ via

$$\mathcal{A}(v_H - \mathcal{Q}v_H, w) = 0 \qquad \text{for all} \quad w \in W, \tag{3.3}$$

where the bilinear form $\mathcal{A}$ is defined as

$$\mathcal{A}(v, \psi) := (\mathfrak{A}(x)\nabla v, \nabla \psi)_\Omega.$$

Due to Assumption 3.1, the linear corrector problem (3.3) has a unique solution. Further, note that (3.3) can equivalently be written as

$$\int_\Omega A_L(x, \nabla u^*, \nabla \mathcal{Q}v_H) \cdot \nabla w \, dx = \int_\Omega A_L(x, \nabla u^*, \nabla v_H) \cdot \nabla w \, dx \qquad \text{for all} \quad w \in W.$$

Having linearized the computation of the correction operator, we define the linear multiscale space $V_H^{\mathrm{ms}} := (\mathrm{id} - \mathcal{Q})V_H$. This problem-adapted space is used in a Galerkin method to seek $u_H^{\mathrm{ms}} \in V_H^{\mathrm{ms}}$ such that

$$\mathcal{B}(u_H^{\mathrm{ms}}; v_H^{\mathrm{ms}}) = (f, v_H^{\mathrm{ms}})_\Omega \qquad \text{for all} \quad v_H^{\mathrm{ms}} \in V_H^{\mathrm{ms}}. \tag{3.4}$$

Let $\{\lambda_z\}_z$ be the nodal basis of V_H (i.e., the standard hat functions). Then $\{\lambda_z - \mathcal{Q}\lambda_z\}_z$ forms a basis of V_H^{ms}. Note that this requires only to solve linear problems. After this basis has been (pre-)computed, the nonlinear problem (3.4) can be solved with any suitable iteration method such as Newton's method, which is rather cheap since the dimension of the multiscale space is small (note $\dim(V_H^{\mathrm{ms}}) = \dim(V_H)$). We emphasize that although we linked nonlinear iterative methods and linearization strategies in Example 3.2, the iterative method chosen to solve (3.4) does not need to correspond to the linearization strategy chosen for the corrector computation. In general, linear problems are much easier and cheaper to solve than nonlinear problems (of the same dimension), which is the appealing advantage of (3.3) in comparison to (3.2) – apart from the discussed well-posedness issues.

3.2. Localization of the basis generation

As in the linear case, the corrector problems (3.3) are global fine-scale problems, which are as expensive to solve as the solution of a (linear) multiscale model problem on a fine-scale mesh resolving the oscillations and discontinuities of A. However since the corrector problems are standard elliptic problems (cf. Assumption 3.1), we can localize these corrector problems in the well-known way for the linear case. To this end, we define the neighborhood

$$\mathrm{N}(T) = \bigcup_{K \in \mathcal{T}_H, T \cap K \neq \emptyset} K$$

associated with an element $T \in \mathcal{T}_H$. Thereby, for any $m \in \mathbb{N}_0$, the m-layer patches are defined inductively via $\mathrm{N}^{m+1}(T) = \mathrm{N}(\mathrm{N}^m(T))$ with $\mathrm{N}^0(T) := T$. The shape regularity implies that there is a bound $C_{\mathrm{ol},m}$ (depending only on m) of the number of the elements in the m-layer patch, i.e.,

$$\max_{T \in \mathcal{T}_H} \mathrm{card}\{K \in \mathcal{T}_H : K \subset \mathrm{N}^m(T)\} \leq C_{\mathrm{ol},m}. \tag{3.5}$$

Throughout this article, we assume that $\mathcal{T}_H$ is quasi-uniform, which implies that $C_{\mathrm{ol},m}$ grows at most polynomially with m.

We then define the truncated correction operator $\mathcal{Q}_m : V_H \to W$ as $\mathcal{Q}_m = \sum_{T \in \mathcal{T}_H} \mathcal{Q}_{T,m}$, where for any $v_H \in V_H$ the truncated element corrector $\mathcal{Q}_{T,m} v_H \in W(\mathrm{N}^m(T)) := \{w \in W : w = 0 \text{ in } \Omega \setminus \mathrm{N}^m(T)\}$ solves

$$\mathcal{A}_{\mathrm{N}^m(T)}(\mathcal{Q}_{T,m} v_H, w) = \mathcal{A}_T(v_H, w) \qquad \text{for all} \quad w \in W(\mathrm{N}^m(T)). \tag{3.6}$$

Here, $\mathcal{A}_D$ denotes the restriction of the bilinear form $\mathcal{A}$ to the subdomain $D \subset \Omega$. With the localized correction operator $\mathcal{Q}_m$, we set up the multiscale space $V_{H,m} := (\mathrm{id} - \mathcal{Q}_m)V_H$. For each element $T \in \mathcal{T}_H$, we only have to solve d problems of type (3.6) with $v_H|_T = x_j$, $j = 1, \ldots, d$, or precisely, the following cell problems: Find $q_{T,m}^{(j)} \in W(\mathrm{N}^m(T))$, $j = 1, \ldots, d$, such that

$$\int_{\mathrm{N}^m(T)} \mathfrak{A}\nabla q_{T,m}^{(j)} \cdot \nabla w \, dx = \int_T \mathfrak{A}e_j \cdot \nabla w \, dx \qquad \text{for all} \quad w \in W(\mathrm{N}^m(T)),$$

where e_j denotes the jth canonical unit vector. Denoting by $\{\lambda_z\}_z$ the standard hat functions, a basis of $V_{H,m}$ is hence given by

$$\left\{\lambda_z - \sum_{T \in \mathcal{T}_H, \, z \in T} \sum_{j=1}^d \left(\frac{\partial}{\partial x_j}\lambda_z|_T\right) q_{T,m}^{(j)}\right\}_z.$$

The localized multiscale method consists of replacing V_H^{ms} by $V_{H,m}$ in (3.4). More precisely, we seek (in a Galerkin method) $u_{H,m} \in V_{H,m}$ such that

$$\mathcal{B}(u_{H,m}; v_{H,m}) = (f, v_{H,m})_\Omega \qquad \text{for all} \quad v_{H,m} \in V_{H,m}. \tag{3.7}$$

As mentioned before, problem (3.7) is solved with an iterative method, e.g., Newton's method, where the multiscale basis can be pre-computed. This requires the storage of all correctors, which can be very memory consuming since $q_{T,m}^{(j)}$ includes fine-scale features. If memory is a limiting factor, the correctors should better be computed on the fly inside each Newton iteration (less memory consuming, but potentially slightly slower). Note that $\mathcal{Q}_m$ and therefore also the solution $u_{H,m}$ depends on $\mathfrak{A}$ and thereby, they implicitly depend on (i) the chosen linearization model and (ii) the chosen "linearization point" u^* in Section 3.1. The choice of u^* and its consequences will be discussed in Section 4.2 below.

Remark 3.3. To avoid communication between the correctors, one can also consider the Petrov-Galerkin method to seek $u_{H,m}^{PG} \in V_H$ such that

$$\mathcal{B}(u_{H,m}^{PG}; v_{H,m}) = (f, v_{H,m})_\Omega \qquad \text{for all} \quad v_{H,m} \in V_{H,m}.$$

In the Petrov-Galerkin method, $q_{T,j}$ and $q_{T',j}$ for $T, T' \in \mathcal{T}_H$ with $T \neq T'$ are never needed at the same time. Hence, these correctors can immediately be discarded once the contributions of element T to the linear system (which is set up in each Newton iteration) are assembled. Hence, when memory becomes the limiting factor, the Petrov-Galerkin variant is highly preferable in comparison to the Galerkin variant, see [15, 16]. Note, however, that $u_{H,m}^{PG}$ only contains information on the coarse scale H. The following error analysis will be restricted to the Galerkin case for simplicity.

Remark 3.4. The present method is still semi-discrete since the corrector problems (3.6) are infinite-dimensional. The discretization procedure for them is equivalent to the case of linear elliptic equations: We introduce a second (fine) simplicial mesh $\mathcal{T}_h$ of Ω which resolves all features of A. Denoting by $V_h = H_0^1(\Omega) \cap \mathcal{S}^1(\mathcal{T}_h)$ the corresponding lowest order Lagrange finite element space, we set $W_h(\mathrm{N}^m(T)) := \{w_h \in V_h : w_h = 0 \text{ in } \Omega \setminus \mathrm{N}^m(T), I_H w_h = 0\}$ and discretize (3.6) by solving over the space $W_h(\mathrm{N}^m(T))$ instead of $W(\mathrm{N}^m(T))$. In the following, we work with the semi-discrete version and emphasize that similar error estimates (with respect to a reference solution $u_h \in V_h$) can be shown in the fully discrete variant, as illustrated for the linear case, see [22, 26, 32].

4. Error analysis

In this section, let the linear model $\mathfrak{A}$ be fixed, i.e., the linearization model A_L and the "linearization point" $u^* \in H_0^1(\Omega)$ are fixed. Occasionally, we will also use $\mathfrak{A}^u := \mathbf{A}_L(\cdot, \nabla u)$ with the above fixed linearization model A_L and the exact solution u to (2.2). $\mathcal{Q}_m^u$ then denotes the corrector associated with $\mathfrak{A}^u$, i.e., $\mathcal{Q}_m^u$ is defined via (3.6) with $\mathfrak{A}$ replaced by $\mathfrak{A}^u$.

Since the linearized corrector problem (3.3) is a standard elliptic problem, we have the following result on the localization error.

Proposition 4.1. *Assume that Assumptions 2.1 and 3.1 are fulfilled. Let $\mathcal{Q}$ be the ideal linearized correction operator defined in (3.3) and $\mathcal{Q}_m$ its truncated/localized version as defined via (3.6). There exists $0 < \beta < 1$ such that for any $v_H \in V_H$*

$$|(\mathcal{Q} - \mathcal{Q}_m)v_H|_1 \lesssim C_{\mathrm{ol},m}^{1/2} \beta^m |v_H|_1.$$

Proposition 4.1 follows from the standard linear elliptic case in [22, 32, 35]. The main idea is that the ideal element corrector $\mathcal{Q}_T$, which is defined via (3.6) with $\mathrm{N}^m(T) = \Omega$, decays exponentially fast (measured in m) away from T. With a slightly different localization strategy, the procedure can also be interpreted in the spirit of an iterative domain decomposition solver, see [29, 30]. Since Assumption 3.1 holds for all $\xi^* \in \mathbb{R}^d$, Proposition 4.1 is still valid if we replace $\mathcal{Q}$ and $\mathcal{Q}_m$ by $\mathcal{Q}^u$ and $\mathcal{Q}_m^u$, respectively.

We first discuss estimates for the Galerkin method (3.7) in Section 4.1. (Additional) error terms arise from the linearization, which we discuss separately in Section 4.2 together with the choice of u^*.

4.1. A priori error estimates

Since $V_{H,m}$ is a linear subspace of $H_0^1(\Omega)$, the Galerkin method (3.4) is automatically well defined, i.e., there exists a unique solution $u_{H,m}$. Its error to the exact solution in the H^1-semi norm can be estimated as follows.

Theorem 4.2. *Let Assumptions 2.1 and 3.1 be fulfilled. Let u be the solution to (2.2) and $u_{H,m}$ the solution to (3.7). Then it holds that*

$$|u - u_{H,m}|_1 \lesssim (H + C_{\mathrm{ol},m}^{1/2} \beta^m)\|f\|_0 + \eta_{\mathrm{lin}}((\mathrm{id} - \mathcal{Q})I_H u) \tag{4.1}$$

as well as

$$|u - u_{H,m}|_1 \lesssim (H + C_{\mathrm{ol},m}^{1/2} \beta^m)\|f\|_0 + \eta_{\mathrm{lin}}(u) \tag{4.2}$$

with, for any $v \in H_0^1(\Omega)$, the linearization error

$$\eta_{\mathrm{lin}}(v) := \sup_{w \in W, |w|_1 = 1} \left| \int_\Omega \left[A(x, \nabla v) - A_L(x, \nabla u^*, \nabla v) \right] \cdot \nabla w \, dx \right|.$$

Further, if $A_L(x, \nabla u, \nabla u) = A(x, \nabla u)$, it also holds that

$$|u - u_{H,m}|_1 \lesssim (H + C_{\mathrm{ol},m}^{1/2} \beta^m)\|f\|_0 + |(\mathcal{Q}_m - \mathcal{Q}_m^u)I_H u|_1. \tag{4.3}$$

Note that the assumption $A_L(x, \nabla u, \nabla u) = A(x, \nabla u)$ is satisfied for the Newton-type and the Kačanov-type linearization introduced in Example 3.2.

Proof of Theorem 4.2. Since (3.7) defines a Galerkin method, the quasi-optimality (2.4) leads to

$$|u - u_{H,m}|_1 \lesssim \inf_{v_{H,m} \in V_{H,m}} |u - v_{H,m}|_1.$$

Proof of (4.1)*:* We choose $v_{H,m} = (\mathrm{id} - \mathcal{Q}_m)I_H u = (\mathrm{id} - \mathcal{Q})I_H u + (\mathcal{Q} - \mathcal{Q}_m)I_H u$ and observe that the second term can directly be estimated using Proposition 4.1, the stability of I_H and (2.3). Note that by definition $u - (\mathrm{id} - \mathcal{Q})I_H u \in W$. Hence we obtain with the strong monotonicity of A (cf. Assumption 2.1), the approximation property (2.6) and (3.3) that

$$
\begin{aligned}
&|u - (\mathrm{id} - \mathcal{Q})I_H u|_1^2 \\
&\lesssim \int_\Omega [A(x, \nabla u) - A(x, \nabla(\mathrm{id} - \mathcal{Q})I_H u)] \cdot \nabla(u - (\mathrm{id} - \mathcal{Q})I_H u) \, dx \\
&= (f, u - (\mathrm{id} - \mathcal{Q})I_H u)_\Omega - \int_\Omega A(x, \nabla(\mathrm{id} - \mathcal{Q})I_H u) \cdot \nabla(u - (\mathrm{id} - \mathcal{Q})I_H u) \, dx \\
&\lesssim (H\|f\|_0 + \eta_{\mathrm{lin}}((\mathrm{id} - \mathcal{Q})I_H u)) |u - (\mathrm{id} - \mathcal{Q})I_H u|_1.
\end{aligned}
$$

9

Combination of this estimate with Proposition 4.1 as described above leads to (4.1).

Proof of (4.2)*:* We again choose $v_{H,m} = (\mathrm{id} - \mathcal{Q}_m)I_H u = (\mathrm{id} - \mathcal{Q})I_H u + (\mathcal{Q} - \mathcal{Q}_m)I_H u$, but we treat the first term in a slightly different manner. Namely, employing Assumption 3.1 and the definition of $\mathcal{Q}$ in (3.3), we deduce

$$
\begin{aligned}
|u - (\mathrm{id} - \mathcal{Q})I_H u|_1^2 &\lesssim \int_\Omega \mathbf{A}_L(x, \nabla u^*)\nabla(u - (\mathrm{id} - \mathcal{Q})I_H u) \cdot \nabla(u - (\mathrm{id} - \mathcal{Q})I_H u)\, dx \\
&= \int_\Omega \big[A_L(x, \nabla u^*, \nabla u) - A_L(x, \nabla u^*, \nabla(\mathrm{id} - \mathcal{Q})I_H u) \big] \cdot \nabla(u - (\mathrm{id} - \mathcal{Q})I_H u)\, dx \\
&= \int_\Omega \big[A_L(x, \nabla u^*, \nabla u) - A(x, \nabla u) + A(x, \nabla u) \big] \cdot \nabla(u - (\mathrm{id} - \mathcal{Q})I_H u)\, dx \\
&= (f, u - (\mathrm{id} - \mathcal{Q})I_H u)_\Omega \\
&\quad + \int_\Omega \big[A_L(x, \nabla u^*, \nabla u) - A(x, \nabla u) \big] \cdot \nabla(u - (\mathrm{id} - \mathcal{Q})I_H u)\, dx \\
&\lesssim (H\|f\|_0 + \eta_{\mathrm{lin}}(u))|u - (\mathrm{id} - \mathcal{Q})I_H u|_1.
\end{aligned}
$$

Combination of this estimate with Proposition 4.1 leads to (4.2).

Proof of (4.3)*:* Again, we choose $v_{H,m} := (\mathrm{id} - \mathcal{Q}_m)I_H u$, but we split it in a different way this time. We write

$$
(\mathrm{id} - \mathcal{Q}_m)I_H u = (\mathrm{id} - \mathcal{Q}^u)I_H u + (\mathcal{Q}^u - \mathcal{Q}_m^u)I_H u + (\mathcal{Q}_m^u - \mathcal{Q}_m)I_H u,
$$

where we recall that $\mathcal{Q}^u$ and $\mathcal{Q}_m^u$ are the solutions to (3.3) and (3.6), respectively, with coefficient $\mathfrak{A}^u := \mathbf{A}_L(x, \nabla u)$. The last term $(\mathcal{Q}_m^u - \mathcal{Q}_m)I_H u$ is directly included in (4.3), while the second term $(\mathcal{Q}^u - \mathcal{Q}_m^u)I_H u$ is again estimated with Proposition 4.1. For the first term $(\mathrm{id} - \mathcal{Q}^u)I_H u$, we observe due to Assumption 3.1 that

$$
\begin{aligned}
|u - (\mathrm{id} - \mathcal{Q}^u)I_H u|_1^2 &\lesssim \int_\Omega \mathbf{A}_L(x, \nabla u)\nabla(u - (\mathrm{id} - \mathcal{Q}^u)I_H u) \cdot \nabla(u - (\mathrm{id} - \mathcal{Q}^u)I_H u)\, dx \\
&= \int_\Omega \big[A_L(x, \nabla u, \nabla u) - A_L(x, \nabla u, \nabla(\mathrm{id}{-}\mathcal{Q}^u)I_H u) \big] \cdot \nabla(u - (\mathrm{id}{-}\mathcal{Q}^u)I_H u)\, dx.
\end{aligned}
$$

Since $u - (\mathrm{id} - \mathcal{Q}^u)I_H u \in W$, the definition of $\mathcal{Q}^u$ implies that

$$
\int_\Omega A_L(x, \nabla u, \nabla(\mathrm{id} - \mathcal{Q}^u)I_H u) \cdot \nabla(u - (\mathrm{id} - \mathcal{Q}^u)I_H u)\, dx = 0.
$$

Hence, employing the assumption $A_L(x, \nabla u, \nabla u) = A(x, \nabla u)$ and (2.6), we deduce

$$
\begin{aligned}
|u - (\mathrm{id} - \mathcal{Q}^u)I_H u|_1^2 &\lesssim \int_\Omega A_L(x, \nabla u, \nabla u) \cdot \nabla(u - (\mathrm{id} - \mathcal{Q}^u)I_H u)\, dx \\
&= \int_\Omega A(x, \nabla u) \cdot \nabla(u - (\mathrm{id} - \mathcal{Q}^u)I_H u)\, dx \\
&= (f, u - (\mathrm{id} - \mathcal{Q}^u)I_H u)_\Omega \\
&\lesssim H\|f\|_0\, |u - (\mathrm{id} - \mathcal{Q}^u)I_H u|_1.
\end{aligned}
$$

This finishes the proof. $\qquad\square$

Up to the linearization errors, which are discussed in Section 4.2, all variants of the previous theorem are identical to the linear elliptic case [32]. In particular, if we choose $m \approx |\log(H)|$, we

have linear convergence of the approximate solution to the exact one without any assumptions on the regularity of u or the variations of A. By Friedrich's inequality, the same estimate also holds for the L^2-norm. In contrast to the linear case, the Aubin-Nitsche trick cannot be applied so that higher order convergence for nonlinear problems is rather difficult to achieve, see the discussion in [2]. Using the idea of the elliptic projection in [2], we obtain an L^2-estimate in Theorem A.1 in the Appendix. Roughly speaking, it yields quadratic convergence (up to (new) linearization errors) for the choice $m \approx |\log(H)|$.

By the stability of I_H we deduce an estimate for the error to $I_H u_{H,m}$, which describes the finite element part of the Galerkin solution.

Corollary 4.3. *Let Assumptions 2.1 and 3.1 be fulfilled. Let u be the solution to (2.2) and $u_{H,m}$ the solution to (3.7). Then it holds that*

$$\|u - I_H u_{H,m}\|_0 \lesssim H \inf_{v_H \in V_H} |u - v_H|_1 + \|u - u_{H,m}\|_0 + H|u - u_{H,m}|_1. \tag{4.4}$$

Proof. With the triangle inequality we split

$$\|u - I_H u_{H,m}\|_0 \leq \|u - I_H u\|_0 + \|I_H(u - u_{H,m})\|_0,$$

which finishes the proof together with the properties (stability, approximation, and projection) of I_H. $\qquad\square$

Note that the two last terms in (4.4) can be estimated via Theorem 4.2. For the L^2-norm we also have the estimates from Theorem A.1 in the appendix. For $m \approx |\log(H)|$, these terms are at least of order H (up to linearization errors) and may even be of order H^2 (cf. the discussion in the appendix). Hence, the error $\|u - I_H u_{H,m}\|_0$ is at least of order H and might be up to order H^2 if u is sufficiently regular. If we assume L^2-stability of I_H, i.e., $\|I_H v\|_0 \lesssim \|v\|_0$ for all $v \in H_0^1(\Omega)$, the estimate in Corollary 4.3 simplifies to

$$\|u - I_H u_{H,m}\|_0 \lesssim \inf_{v_H \in V_H} \|u - v_H\|_0 + \|u - u_{H,m}\|_0.$$

Together with Theorem A.1 this implies that the error in the finite element part of $u_{H,m}$ is dominated by the L^2-best-approximation error in the finite element space.

4.2. Linearization errors and choice of linearization points

In this section, we discuss estimates for the linearization errors of Theorem 4.2, namely $\eta_{\mathrm{lin}}(v)$ and $|(\mathcal{Q}_m - \mathcal{Q}_m^u) I_H u|_1$, and their implication on the choice of the "linearization point" u^*.

Linearization error $\eta_{\mathrm{lin}}(v)$. First, we note that due to Assumptions 2.1 and 3.1, $\eta_{\mathrm{lin}}(v)$ can always be bounded as follows

$$
\begin{aligned}
\eta_{\mathrm{lin}}(v) &\leq \|A(x, \nabla v) - A_L(x, \nabla u^*, \nabla v)\|_0 \\
&\leq \|A(x, \nabla v) - A(x, 0)\|_0 + \|\mathbf{A}_L(x, \nabla u^*) \nabla v\|_0 + \|A(x, 0)\|_0 + \|b_L(x, \nabla u^*)\|_0 \\
&\lesssim (\Lambda + \Lambda_L)|v|_1 + C_0 + C_b.
\end{aligned}
$$

Hence, the stabilities of $\mathcal{Q}$ and I_H as well as (2.3) directly imply that $\eta_{\mathrm{lin}}(u) \leq C\|f\|_0 + C$ and $\eta_{\mathrm{lin}}((\mathrm{id} - \mathcal{Q})I_H u) \leq C\|f\|_0 + C$, i.e., this linearization error is bounded from above without any further assumptions. As a consequence, if the nonlinearity and the right-hand side data are small, the linearization error is negligible and the discretization error is dominating in (4.1) and (4.2). This of course is a rather restrictive assumption since it basically means that we are still

11

in the almost linear case with only a small nonlinear perturbation. Note that the above bound on $\eta_{\lin}(v)$ can be simplified in the case $v = (\id - \mathcal{Q})I_H u$ because due to (3.3)

$$\int_\Omega A_L(x, \nabla u^*, (\id - \mathcal{Q})I_H u) \cdot \nabla w \, dx = 0 \qquad \text{for all} \quad w \in W.$$

Next, we will show that $\eta_{\lin}(v)$ is small for the linearizations of Example 3.2 if u^* is close to the exact solution u.

Lemma 4.4 (Linearization error for Kačanov-type linearization). *Let* $A(x, \xi) = \alpha(x, |\xi|^2)\xi$ *and set* $\mathbf{A}_L(x, \xi^*) = \alpha(x, |\xi^*|^2)$ *and* $b_L(x, \xi^*) = 0$. *Then,*

$$\eta_{\lin}(v) \leq (\Lambda + \Lambda_L)|v - u^*|_1.$$

Proof. Assumptions 2.1 and 3.1 directly yield

$$
\begin{aligned}
\eta_{\lin}(v) &\leq \|\alpha(x, |\nabla v|^2)\nabla v - \alpha(x, |\nabla u^*|^2)\nabla v\|_0 \\
&\leq \|A(x, \nabla v) - A(x, \nabla u^*)\|_0 + \|\alpha(x, |\nabla u^*|^2)\nabla(u^* - v)\|_0 \\
&\leq (\Lambda + \Lambda_L)\|\nabla(v - u^*)\|_0,
\end{aligned}
$$

which finishes the proof. $\qquad\square$

Lemma 4.5 (Linearization error for Newton-type linearization). *Set* $\mathbf{A}_L(x, \xi^*) = D_\xi A(x, \xi^*)$ *and* $b_L(x, \xi^*) = A(x, \xi^*) - D_\xi A(x, \xi^*)\xi^*$. *Assume that* $D_\xi A$ *is Lipschitz continuous in its last argument, i.e., there is* $L_A > 0$ *such that*

$$|D_\xi A(x, \xi_1) - D_\xi A(x, \xi_2)| \leq L_A|\xi_1 - \xi_2| \qquad \text{for all} \quad \xi_1, \xi_2 \in \mathbb{R}^d \quad \text{and almost all} \quad x \in \Omega. \quad (4.5)$$

Then,

$$\eta_{\lin}(v) \leq L_A \|\nabla(v - u^*)\|_{L^\infty(\Omega)} |v - u^*|_1.$$

Proof. We perform a Taylor expansion of A around u^* and obtain that for any $w \in H_0^1(\Omega)$ it holds that

$$\int_\Omega \left[A(x, \nabla v) - A(x, \nabla u^*) + D_\xi A(x, \nabla u^*)\nabla(v - u^*) \right] \cdot \nabla w \, dx$$

$$= \int_\Omega \int_0^1 \left[D_\xi A(x, \nabla u^* + \tau \nabla(v - u^*)) - D_\xi A(x, \nabla u^*) \right] \nabla(v - u^*) \cdot \nabla w \, d\tau \, dx.$$

The assumed Lipschitz continuity of $D_\xi A$ (4.5) finishes the proof. $\qquad\square$

Note that we did not exploit $w \in W$ in the estimates for $\eta_{\lin}(v)$. Lemmas 4.4 and 4.5 underline that u^* should be close to u, which can be seen as a local convergence result and motivates the discussion of different linearization points further below in this section.

Linearization error $|(\mathcal{Q}_m - \mathcal{Q}_m^u)I_H u|_1$. In case of a Kačanoc-type linearization from Example 3.2, we observe that $\eta_{\lin}(u)$ in (4.2) can be bounded as $\eta_{\lin}(u) \leq \|\mathfrak{A}^u - \mathfrak{A}\|_{L^\infty(\Omega)}|u|_1$. Hence, this linearization error can be expressed as the $L^\infty(\Omega)$-error between the "true" coefficient $\mathfrak{A}^u$ and the "reference" coefficient $\mathfrak{A}$ used in practice for computing the corrector $\mathcal{Q}_m$. The error $|(\mathcal{Q}_m - \mathcal{Q}_m^u)I_H u|_1$ in (4.3) is obviously also closely related to the error between the coefficients $\mathfrak{A}^u$ and $\mathfrak{A}$. However, estimating directly the error between the correction operators allows for a refined estimate in the sense that not merely the global L^∞-error, which is of the order one

already for a local change in the coefficient, will be taken into account. Before we present the estimate in detail, let us first note that, obviously,

$$|(\mathcal{Q}_m - \mathcal{Q}_m^u)I_H u|_1 \le \sup_{v_H \in V_H, |v_H|_1 = 1} |(\mathcal{Q}_m - \mathcal{Q}_m^u)v_H|_1 \, |I_H u|_1$$

so that we will estimate $(\mathcal{Q}_m - \mathcal{Q}_m^u)v_H$ for an arbitrary $v_H \in V_H$ in the following. Further, by the definition of $\mathcal{Q}_m$ and $\mathcal{Q}_m^u$ as sums of element correctors with support only in $\mathrm{N}^m(T)$ we have

$$|(\mathcal{Q}_m - \mathcal{Q}_m^u)v_H|_1^2 \lesssim C_{\mathrm{ol},m} \sum_{T \in \mathcal{T}_H} |(\mathcal{Q}_{T,m} - \mathcal{Q}_{T,m}^u)v_H|_1^2.$$

We now have the following estimate for the error between the correction operators, the proof is postponed to the Appendix B.

Proposition 4.6. *Fix $T \in \mathcal{T}_H$ and $v_H \in V_H$. For any $a \in L^\infty(\mathrm{N}^m(T), \mathbb{R}^{d \times d})$ denote by $\bar{a}$ the average of the trace, i.e., $\bar{a} := \fint_{\mathrm{N}^m(T)} \mathrm{tr}(a)\,dx$, and denote by $\hat{a} \in L^\infty(\mathrm{N}^m(T), \mathbb{R}^{d \times d})$ the scaled coefficient, i.e., $\hat{a} := a/\bar{a}$. Define*

$$E_{\mathcal{Q},T}^2 := \sum_{T' \in \mathcal{T}_H, T' \subset \mathrm{N}^m(T)} \|\widehat{\mathfrak{A}^u} - \widehat{\mathfrak{A}}\|_{L^\infty(T')}^2 \max_{\psi|_T, \psi \in V_H} \frac{\|(\chi_T \nabla\psi - \nabla\mathcal{Q}_{T,m}\psi)\|_{0,T'}^2}{|\psi|_{1,T}^2}.$$

Then,

$$|(\mathcal{Q}_{T,m} - \mathcal{Q}_{T,m}^u)v_H|_1 \lesssim E_{\mathcal{Q},T}|v_H|_{1,T}.$$

We emphasize that Proposition 4.6 relates the error between the element correctors to the error between the coefficients $\mathfrak{A}$ and $\mathfrak{A}^u$, but (i) locally on each element or element patch, respectively, and (ii) only the L^∞-error between the *scaled* coefficients is relevant. The latter point is possible because (3.6) can be multiplied by the scalar-valued constant $\overline{\mathfrak{A}}$ without changing the element corrector. Note that this is only possible for scalar-valued constants which is the reason for taking the trace of A in the definition of $\bar{a}$. Hence, if a coefficient is (locally) multiplied by a constant, the error between the associated correction operators is zero while the error between the unscaled coefficients can become very large. This is the second advantage we see in estimating the error between the correction operators.

Error indicators similar to $E_{\mathcal{Q},T}$ in Proposition 4.6 were already presented in [19, 20] in the context of time-changing or perturbed diffusion coefficients. We point out the following differences between Proposition 4.6 and the results in [19, 20]: First, the idea of studying the scaled coefficients is new in Proposition 4.6. Second, since $\mathfrak{A}^u$ is not available in practice in our case, Proposition 4.6 is an a priori result and aims at linking the error between the correction operators to the error between the coefficients. This is in sharp contrast to the previous works, which use $E_{\mathcal{Q},T}$ as a practical indicator to steer re-computation of element correctors. For this reason, some terms in [19, 20] are different or additional to (i) estimate the error between the element correctors in the energy norm and (ii) to avoid constants depending on the upper and lower spectral bounds of the diffusion coefficients in the error estimate. The result of Proposition 4.6 with the scaled coefficients can probably be refined in this spirit as well, but we omit this for simplicity.

Choice of linearization points u^*. The previous discussion of the linearization errors has shown that the choice of u^* is crucial for the performance of the method and that, ideally, u^* should be chosen close the exact, but unknown, solution u. With this in mind, let us discuss and compare some possible choices of u^*.

1. We can select for u^* the initial value of the nonlinear iteration, for instance $u^* = 0$. This results in a multiscale space $V_{H,m}$ that can be computed a priori. Since $V_{H,m}$ is rather low-dimensional, this is a very cheap method. The overall error of course will only be dominated by the discretization error of order $H + \beta^m$ if u^* is already close to the exact solution. For instance, a similar assumption is also needed to guarantee convergence of Newton's iteration method for the nonlinear problem. In particular the choice $u^* = 0$ should yield satisfactory results for problems without steep gradients. We discuss this issue also in the numerical examples.

2. Since choosing u^* a priori will work only in rather restrictive settings, another option is to compute the finite element solution u_H using V_H and to use it as u^*. Since V_H is associated with a coarse mesh, the computation of u_H is rather cheap. Due to the finescale features of A, however, u_H alone will not be a satisfactory approximation of u in general, not even in the $L^2(\Omega)$-norm. The finescale features of A are then taken into account in the ensuing LOD solution. To estimate the linearization error in this case, we can combine Lemmas 4.4 or 4.5, respectively, with standard a priori estimates for the finite element solution. Finite element error estimates in the $L^2(\Omega)$-norm and $H^1(\Omega)$-semi norm are briefly presented in Section 2.2. In case of the Newton-type linearization, Lemma 4.5 also requires a $W^{1,\infty}(\Omega)$-estimate, where we obtain from [8, Chapter 8]

$$\|u - u_H\|_{W^{1,\infty}(\Omega)} \leq CH\|u\|_{W^{2,\infty}(\Omega)}.$$

Note that for all finite element error estimates, higher regularity of u is required and that the constants will depend on the finescale parameter ε, cf. the discussion in Section 2.2. We emphasize that this approach is different from [39] where u_H is "post-processed" by solving a linear problem on a fine mesh.

3. Finally, one can compute a whole cascade of LOD solutions by starting with some u^*, computing the solution $u_{H,m}$ to (3.7), and using this solution or its FE part $I_H u_{H,m}$ as new u^*. For this cascade, we can inductively combine (4.2) and Lemmas 4.4 or 4.5, respectively, to obtain an a priori error estimate, which consists of the discretization error of order $H + \beta^m$ and the linearization error for the initial choice of u^*. This straightforward procedure of estimating the error, however, does not seem to be optimal because the initial linearization error will always remain and one does not exploit that the LOD solution should be closer to the exact solution with each step.

   When computing a whole cascade of LOD solutions by a nonlinear solve of (3.7) and an update of the linearization point, an error indicator in the spirit of Proposition 4.6 – with $\mathcal{Q}_m^u$ replaced by $\mathcal{Q}_m^{u_H,m}$ with the current LOD solution – can be used to locally determine which element correctors should be recomputed with the new linearization point. Since (3.7) is solved by an iterative method as well, one can also imagine to already update some element correctors during this nonlinear iteration. A detailed study of such iterative generalized finite element approaches – for instance concerning convergence of the iteration, suitable stopping criteria and a priori error analysis – is a future research topic of its own and beyond the scope of the present contribution.

In the numerical experiments, we will compare the solution to (3.7) with (i) $u^* = 0$ as in the first point, (ii) $u^* = u_H$ as in the second point, (iii) $u^* = u_{H,m}$ or $u^* = I_H u_{H,m}$ with $u_{H,m}$ the LOD solution with linearization point zero as in the third point.
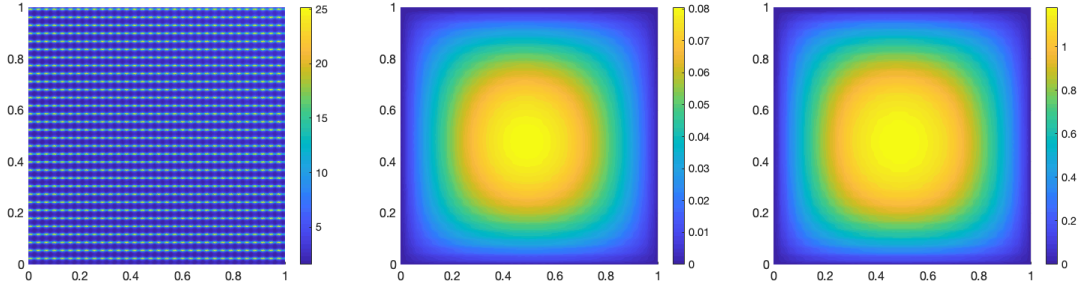
Figure 5.1: Spatial part of the coefficient (left) and reference solutions u_h for f_1(middle) and f_2 (right) in the experiment of Section 5.1.

5. Numerical experiments

We present the results of several numerical experiments, subject to different multiscale coefficients and nonlinearities. In all cases, the computational domain is $\Omega = [0,1]^2$. Since no exact solution is known in any of the examples, we compute a reference solution $u_h \in V_h$ using a standard finite element method on a fine mesh $\mathcal{T}_h$ that resolves all multiscale features. Specifically, the reference mesh size is fixed as $h = 2^{-8}$ in all experiments and the coarse mesh size varies as $H = 2^{-2}, 2^{-3}, \ldots 2^{-6}$. We present results for oversampling parameters $m = 1, 2, 3$ and already point out that $m = 2, 3$ is a sufficient choice in most experiments. To solve the nonlinear problems, we use Newton's method with tolerance 10^{-11} for the residual as stopping criterion. Sections 5.1 and 5.2 consider nonlinearities of the type $A(x, \nabla u)$, where Assumption 2.1 is only satisfied in Section 5.1. Further, we consider a model for the stationary Richards equation with a quasilinear coefficient of the form $a(x, u)\nabla u$ in Section 5.3. Note that the resulting nonlinear form $\mathcal{B}$ is no longer monotone such that the above proof techniques do not directly transfer, see [5, 6]. Nevertheless, we use the presented multiscale method with the obvious modifications for (3.7) and $\mathfrak{A} := a(x, 0)$ in the corrector problems.

5.1. Periodic coefficient

We choose a model problem similar to [2]. The nonlinear coefficient is defined as

$$A(x, \xi) = \left(1 + x_1 x_2 + \frac{1.1 + \frac{\pi}{3} + \sin(2\pi \frac{x}{\varepsilon})}{1.1 + \sin(2\pi \frac{x}{\varepsilon})}\right)\left(1 + \frac{1}{\left(1 + |\xi|^2\right)^{1/2}}\right)\xi$$

with $\varepsilon = 2^{-5}$ and sources $f_1(x) = 10\exp(-0.1|x - x_0|^2)$ or $f_2(x) = 100\exp(-0.1|x - x_0|^2)$ with $x_0 = (0.45, 0.5)^T$ in our case. The coefficient and the reference solutions u_h are depicted in Figure 5.1. We will consider the two right-hand sides f_1 and f_2 to study the influence of higher values of the solution u and its gradient on our method and the linearization errors. Note that one can hope for higher regularity of the exact solution u in this experiment.

We use the Galerkin method (3.7) and obtain a solution $u_{H,m} \in V_{H,m}$, where the correctors are computed using the Newton-type linearization at u^*, which is specified below. We focus on two (relative) errors in the following: The so-called *(relative) upscaled error*

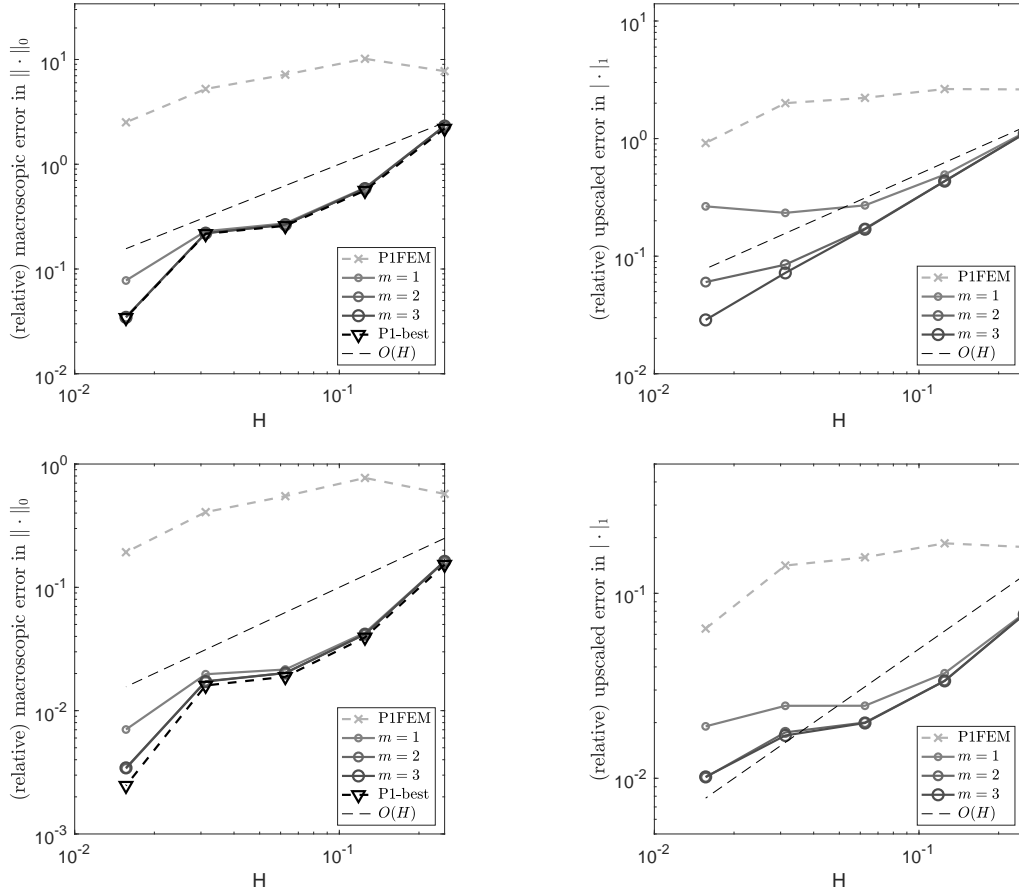$$e_{\mathrm{LOD}} := \frac{|u_h - u_{H,m}|_1}{|u_h|_1}$$

Figure 5.2: Convergence histories for the (relative) macroscopic error e_H (in L^2-norm) (left) and the (relative) upscaled error e_{LOD} (in H^1-semi norm) (right) with f_1 (top) and f_2 (bottom) for the experiment of Section 5.1.

between the reference and the LOD solution in the H^1-semi norm, for which we expect a linear convergence rate (cf. Theorem 4.2); and the so-called *(relative) macroscopic error*

$$e_H := \frac{\|u_h - I_H u_{H,m}\|_0}{\|u_h\|_0}$$

between the reference and the FE part of the LOD solution in the L^2-norm, for which we expect the same behavior as the L^2-best approximation in V_H (cf. Theorem 4.3).

For the simple choice $u^* = 0$, these two errors are depicted for the two right-hand sides in Figure 5.2. We note that the (relative) macroscopic errors e_H in the left column closely follow the error of the (relative) L^2-best approximation in the space V_H for $m = 2, 3$ (cf. the discussion after Corollary 4.3). Since $I_H u_{H,m}$ lies in the same space, we cannot hope for anything better. In particular, the reduced convergence rate for H between $\sqrt{\varepsilon}$ and ε is no defect of the method, but intrinsic to the problem, see also the discussion of this so-called resonance effect in [18]. We emphasize that in the pre-asymptotic range $H \ll \varepsilon$, the standard finite element method does not produce faithful results and also shows no convergence rates, while the presented multiscale
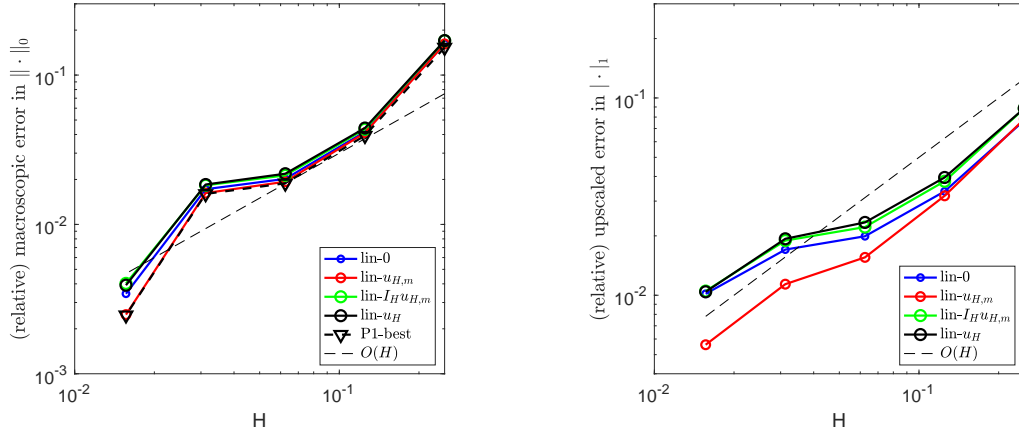
16

Figure 5.3: Convergence histories for the (relative) macroscopic error e_H (in L^2-norm) (left) and the (relative) upscaled error e_{LOD} (in H^1-semi norm) (right) with f_2 and different choices of the linearization point for the experiment of Section 5.1.

method follows the best approximation error and converges with linear rate. For approximations in the H^1-semi norm, the (coarse-scale) space V_H is no longer sufficient, as also underlined by homogenization theory, but fine-scale features have to be taken in account. Therefore, we consider the (relative) upscaled error e_{LOD} in the right column of Figure 5.2. This error for the LOD solution $u_{H,m}$ overall converges linearly as expected from Theorem 4.2. All in all, the experiment clearly confirms the predicted convergence rates of Theorem 4.2 and Corollary 4.3.

When comparing the top and bottom row of Figure 5.2, we also observe that the higher values of u_h and thereby larger gradients of u_h caused by f_2 influence the performance of the method. In particular the relative upscaled error e_{LOD} in H^1-semi norm shows some deviation from the optimal linear convergence most probably because the linearization error starts to dominate. Therefore, we compare the behavior of e_H and e_{LOD} for different linearization points with fixed $m = 3$ in Figure 5.3. We consider the following choices of u^* as discussed in Section 4.2: $u^* = 0$, $u^* = u_H$ with the standard FE solution u_H on the coarse mesh $\mathcal{T}_H$ and $u^* = u_{H,m}$ as well as $u^* = I_H u_{H,m}$, where $u_{H,m}$ is the LOD solution with linearization at zero. All these solutions show a similar qualitative behavior, but in the quantitative errors we observe clear differences between the linearization points, especially for e_{LOD}. The choices $u^* = u_H$ and $u^* = I_H u_{H,m}$ result in almost identical results, probably because they lie in the same space V_H. A bit surprisingly, those choices even perform slightly worse than the simple choice $u^* = 0$. A possible explanation is that while $I_H u_{H,m}$ is already a good approximation of u in the L^2-sense, the gradients are relevant for the nonlinearity, where the space V_H does not provide sufficient approximations. Both for e_H and e_{LOD}, the choice $u^* = u_{H,m}$ performs best, which motivates to study iterative LOD approximations as discussed in Section 4.2 in more detail in future research.

5.2. Random coefficient

We choose the nonlinear coefficient as

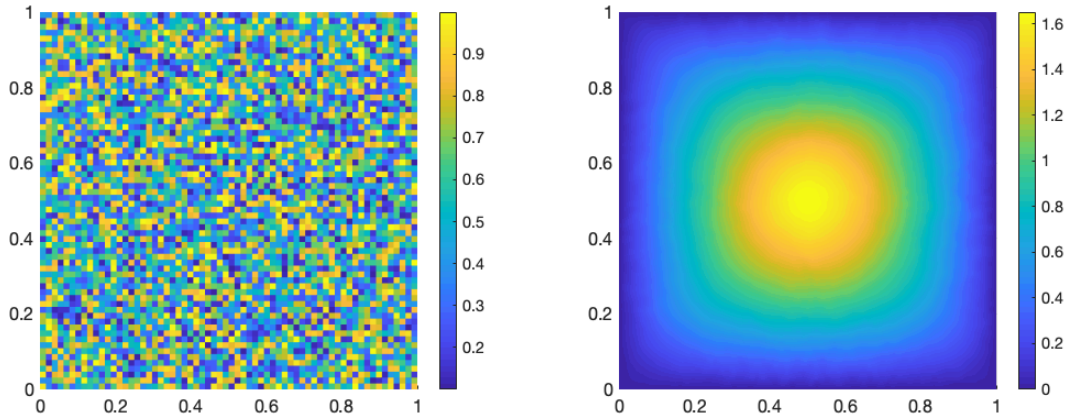$$A(x, \xi) = c(x) \begin{pmatrix} \xi_1 + \frac{1}{3}\xi_1^3 \\ \xi_2 + \frac{1}{3}\xi_2^3 \end{pmatrix},$$

17

Figure 5.4: Spatial part of the coefficient (left) and reference solution u_h (right) for the experiment of Section 5.2.

where $c(x)$ is piece-wise constant on a quadrilateral mesh $\mathcal{T}_\varepsilon$ with $\varepsilon = 2^{-6}$ and the values are random numbers in $[0.1, 1]$. The right-hand side is

$$f(x) = \begin{cases} 5 & x_2 \leq 0.1, \\ 50 & \text{else,} \end{cases}$$

see [21] for the nonlinearity and a similar right-hand side. Note that we clearly cannot expect higher regularity than $H_0^1(\Omega)$ for the exact solution in this case because of the spatial discontinuities in A. Further, A is only locally Lipschitz constant in its second argument. Thus, Assumption 2.1 is violated and we even have $|A| \to \infty$ for $|\xi| \to \infty$. The coefficient and corresponding reference solution u_h are depicted in Figure 5.4.

For this example, we study the Petrov-Galerkin LOD as briefly discussed in Remark 3.3. Precisely, we compute $u_{H,m}^{PG} \in V_H$ as the solution of

$$\mathcal{B}(u_{H,m}^{PG}; v_H) = (f, v_H)_\Omega \qquad \text{for all} \quad v_H \in V_{H,m},$$

where $V_{H,m}$ is defined as in the Galerkin case. Hence, the solution $u_{H,m}^{PG}$ lies in the FE space and we expect the relative macroscopic error $e_H := \frac{\|u_h - u_{H,m}\|_0}{\|u_h\|_0}$ to follow the L^2-best approximation error as predicted in Corollary 4.3 for the FE part of the (Galerkin) LOD solution.

We first consider the convergence history of e_H for the linearization $u^* = 0$ and different choices of m in Figure 5.5 (left). The multiscale method performs obviously better than the standard FEM. The error of the Petrov-Galerkin LOD is following the L^2-best approximation as expected up to a saturation or stagnation for the last considered mesh, where most probably the linearization error starts to dominate. As in the previous example, we also consider e_H for fixed m and different linearization points u^* in Figure 5.5 (right). The considered choices are $u^* = 0$, $u^* = u_H$ with the FE solution u_H on the coarse mesh as before and $u^* = u_{H,m}^{PG}$ as well as $u^* = (\mathrm{id} - \mathcal{Q}_m) u_{H,m}^{PG}$, where $u_{H,m}^{PG}$ is the Petrov-Galerkin LOD solution with $u^* = 0$ and $\mathcal{Q}_m$ is the corrector computed with $u^* = 0$. The latter two linearization points play roles comparable to $I_H u_{H,m}$ and $u_{H,m}$ in the Galerkin setting discussed in the previous experiment. Although the linearization with zero shows a saturation at the end, it performs best of the considered
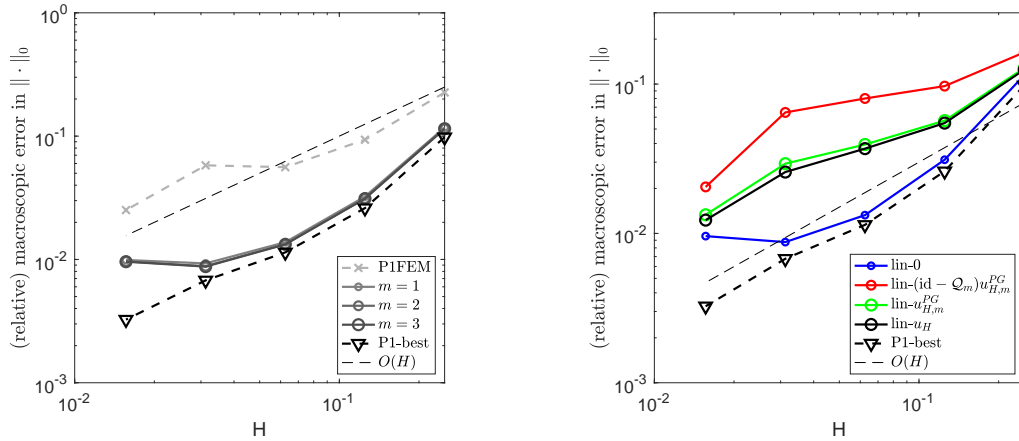
Figure 5.5: Convergence histories for the (relative) macroscopic error e_H (in L^2-norm) for fixed $u^* = 0$ and different m (left) and for fixed $m = 3$ and different u^* (right) for the experiment of Section 5.2.

linearization points. We emphasize that the impractical choice of $u^* = u_h$ would lead to an e_H completely following the L^2-best approximation error (data not shown). This illustrates the validity of our error estimates also for the Petrov-Galerkin LOD, but also stresses the influence of the chosen linearization point. Moreover, this example confirms and underlines that the present multiscale method does not rely on assumptions such as periodicity or scale separation and thus also works in situations where no homogenization results are available.

5.3. Stationary Richards equation

We now test the applicability of our method to quasilinear non-monotone problems, which, for instance, are frequently encountered in (unsaturated) groundwater flow and can be modeled by the (stationary) Richards equation. Here, we consider a quasilinear coefficient of the form $A(x, u, \nabla u) = c(x)k(u)\nabla u$. For c we choose a spatial multiscale model with a channel as depicted in Figure 5.6 left, which is often present in geophysical applications. Note that this can easily be extended to the case of several channels. For the nonlinearity k, we consider the following so-called van Genuchten model [38]

$$k(s) = \frac{(1 - \alpha|s|(1 + (\alpha|s|)^2)^{-1/2})^2}{1 + (\alpha|s|)^2}$$

with $\alpha = 0.005$, see also [11]. The right-hand side is

$$f(x) = \begin{cases} 0.1 & x_2 \leq 0.1, \\ 1 & \text{else}, \end{cases}$$

The reference solution for the present setting is depicted in Figure 5.6 right. Note the influence of the channel, which analytically manifests itself in a low regularity of the solution.

As a consequence, we observe exactly the (worst-case) convergence rates as predicted by our theory (for the monotone case only), but not more. In particular, the L^2-error for $I_H u_{H,m}$ follows the best-approximation error, which in this case is "only" linear as discussed after Corollary 4.3.
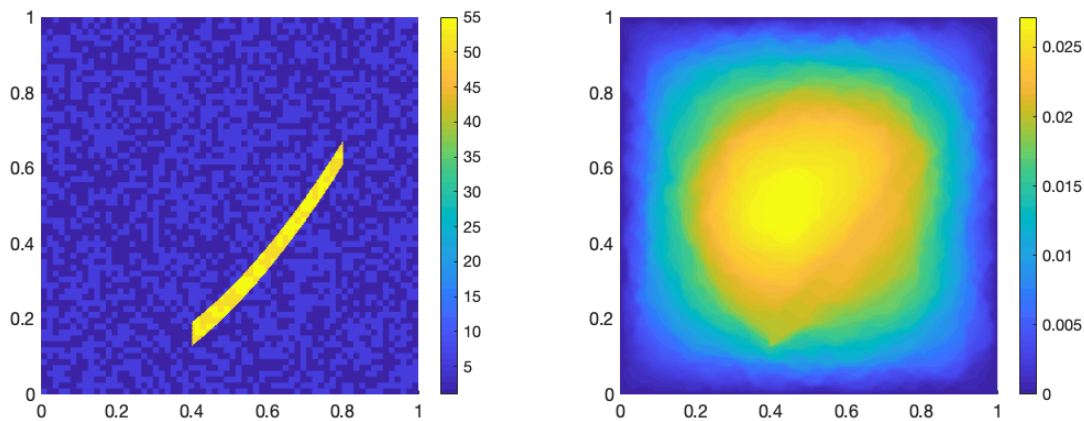
Figure 5.6: Coefficient $c(x)$ (left) and reference solution (right) for the experiment of Section 5.3.

Furthermore, we see the expected linear convergence of the upscaled error e_{LOD} up to a slight saturation for small H. This is most probably caused again by a dominating linearization error, which could be cured by a different choice of the linearization point as discussed in the previous experiments and in Section 4.2. This experiment clearly underlines the applicability of the approach also beyond the strictly monotone case and indicates that similar convergence rates as in Theorem 4.2 and Corollary 4.3 can be expected.

Conclusion

We presented a multiscale method for nonlinear monotone elliptic problems with spatial multiscale features. A problem-adapted multiscale basis is constructed by solving local linear fine-scale problems for each coarse-scale mesh element. Numerical analysis shows optimal error estimates up to linearization errors and we discussed choices of the linearization. Several numerical experiments underline and confirm the applicability of the method as well as the expected convergence rates. We also numerically compared the influence of the chosen linearization on the performance of the method. As mentioned, iterative (adaptive) LOD methods where either a cascade of LOD solutions to different linearization points is computed or the correctors are (partially) updated during the nonlinear iteration are interesting extensions of the presented method and will be subject of future research. Also the numerical analysis for the quasilinear non-monotone problems, where the numerical example has shown promising results, will be studied in the future. A generalization of the method to other nonlinear problems seems possible, here we note as interesting applications (besides Richards equation) porous media flow, nonlinear elasticity, and wave propagation in nonlinear media (e.g., of Kerr-type).

References

[1] A. Abdulle, Y. Bai, and G. Vilmart. Reduced basis finite element heterogeneous multiscale method for quasilinear elliptic homogenization problems. *Discrete Contin. Dyn. Syst. Ser. S*, 8(1):91–118, 2015.

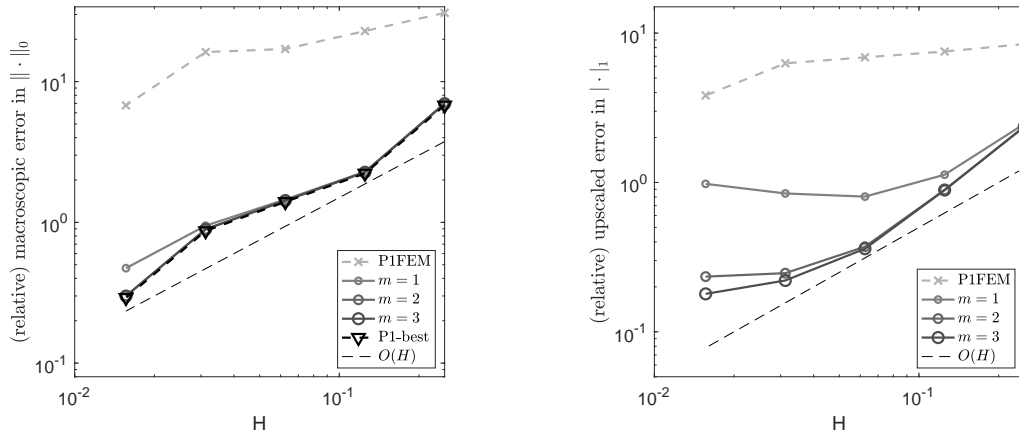[2] A. Abdulle and M. E. Huber. Error estimates for finite element approximations of nonlinear

Figure 5.7: Convergence histories for the (relative) macroscopic error (in L^2-norm) (left) and for the (relative) upscaled error e_{LOD} (in H^1-semi norm) (right) for the experiment of Section 5.3.

monotone elliptic problems with application to numerical homogenization. *Numer. Methods Partial Differential Equations*, 32(3):955–969, 2016.

[3] A. Abdulle and M. E. Huber. Finite element heterogeneous multiscale method for non-linear monotone parabolic homogenization problems. *ESAIM Math. Model. Numer. Anal.*, 50(6):1659–1697, 2016.

[4] A. Abdulle, M. E. Huber, and G. Vilmart. Linearized numerical homogenization method for nonlinear monotone parabolic multiscale problems. *Multiscale Model. Simul.*, 13(3):916–952, 2015.

[5] A. Abdulle and G. Vilmart. A priori error estimates for finite element methods with numerical quadrature for nonmonotone nonlinear elliptic problems. *Numer. Math.*, 121(3):397–431, 2012.

[6] A. Abdulle and G. Vilmart. Analysis of the finite element heterogeneous multiscale method for quasilinear elliptic homogenization problems. *Math. Comp.*, 83(286):513–536, 2014.

[7] G. Allaire. Homogenization and two-scale convergence. *SIAM J. Math. Anal.*, 23(6):1482–1518, 1992.

[8] S. C. Brenner and L. R. Scott. *The mathematical theory of finite element methods*, volume 15 of *Texts in Applied Mathematics*. Springer-Verlag, New York, 1994.

[9] E. Chung, Y. Efendiev, K. Shi, and S. Ye. A multiscale model reduction method for nonlinear monotone elliptic equations in heterogeneous media. *Netw. Heterog. Media*, 12(4):619–642, 2017.

[10] P. G. Ciarlet. *The finite element method for elliptic problems*, volume 40 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002.

[11] Y. Efendiev, J. Galvis, S. Ki Kang, and R. D. Lazarov. Robust multiscale iterative solvers for nonlinear flows in highly heterogeneous media. *Numer. Math. Theory Methods Appl.*, 5(3):359–383, 2012.

[12] Y. Efendiev, J. Galvis, G. Li, and M. Presho. Generalized multiscale finite element methods. Nonlinear elliptic equations. *Commun. Comput. Phys.*, 15(3):733–755, 2014.

[13] Y. Efendiev, T. Y. Hou, and V. Ginting. Multiscale finite element methods for nonlinear problems and their applications. *Commun. Math. Sci.*, 2(4):553–589, 2004.

[14] L. El Alaoui, A. Ern, and M. Vohralík. Guaranteed and robust a posteriori error estimates and balancing discretization and linearization errors for monotone nonlinear problems. *Comput. Methods Appl. Mech. Engrg.*, 200(37-40):2782–2795, 2011.

[15] D. Elfverson, V. Ginting, and P. Henning. On multiscale methods in Petrov-Galerkin formulation. *Numer. Math.*, 131(4):643–682, 2015.

[16] C. Engwer, P. Henning, A. Målqvist, and D. Peterseim. Efficient implementation of the localized orthogonal decomposition method. *Comput. Methods Appl. Mech. Engrg.*, 350:123–153, 2019.

[17] M. Feistauer and A. Ženíšek. Finite element solution of nonlinear elliptic problems. *Numer. Math.*, 50(4):451–475, 1987.

[18] D. Gallistl and D. Peterseim. Computation of quasi-local effective diffusion tensors and connections to the mathematical theory of homogenization. *Multiscale Model. Simul.*, 15(4):1530–1552, 2017.

[19] F. Hellman, T. Keil, and A. Målqvist. Numerical upscaling of perturbed diffusion problems. *SIAM J. Sci. Comput.*, 42(4):A2014–A2036, 2020.

[20] F. Hellman and A. Målqvist. Numerical homogenization of elliptic PDEs with similar coefficients. *Multiscale Model. Simul.*, 17(2):650–674, 2019.

[21] P. Henning. *Heterogeneous multiscale finite element methods for advection-diffusion and nonlinear elliptic multiscale problems.* PhD thesis, WWU Münster, 2011.

[22] P. Henning and A. Målqvist. Localized orthogonal decomposition techniques for boundary value problems. *SIAM J. Sci. Comput.*, 36(4):A1609–A1634, 2014.

[23] P. Henning, A. Målqvist, and D. Peterseim. A localized orthogonal decomposition method for semi-linear elliptic problems. *ESAIM Math. Model. Numer. Anal.*, 48(5):1331–1349, 2014.

[24] P. Henning, A. Målqvist, and D. Peterseim. Two-level discretization techniques for ground state computations of Bose-Einstein condensates. *SIAM J. Numer. Anal.*, 52(4):1525–1550, 2014.

[25] P. Henning and M. Ohlberger. Error control and adaptivity for heterogeneous multiscale approximations of nonlinear monotone problems. *Discrete Contin. Dyn. Syst. Ser. S*, 8(1):119–150, 2015.

[26] P. Henning and D. Peterseim. Oversampling for the multiscale finite element method. *Multiscale Model. Simul.*, 11(4):1149–1175, 2013.

[27] V. H. Hoang. Sparse finite element method for periodic multiscale nonlinear monotone problems. *Multiscale Model. Simul.*, 7(3):1042–1072, 2008.

[28] M. Huber. *Numerical homogenization methods for advection-diffusion and nonlinear monotone problems with multiple scales.* PhD thesis, EPFL, 2015.

[29] R. Kornhuber, D. Peterseim, and H. Yserentant. An analysis of a class of variational multiscale methods based on subspace decomposition. *Math. Comp.*, 87(314):2765–2774, 2018.

[30] R. Kornhuber and H. Yserentant. Numerical homogenization of elliptic multiscale problems by subspace decomposition. *Multiscale Model. Simul.*, 14(3):1017–1036, 2016.

[31] D. Lukkassen, G. Nguetseng, and P. Wall. Two-scale convergence. *Int. J. Pure Appl. Math.*, 2(1):35–86, 2002.

[32] A. Målqvist and D. Peterseim. Localization of elliptic multiscale problems. *Math. Comp.*, 83(290):2583–2603, 2014.

[33] H. Owhadi. Multigrid with rough coefficients and multiresolution operator decomposition from hierarchical information games. *SIAM Rev.*, 59(1):99–149, 2017.

[34] H. Owhadi and L. Zhang. Localized bases for finite-dimensional homogenization approximations with nonseparated scales and high contrast. *Multiscale Model. Simul.*, 9(4):1373–1398, 2011.

[35] D. Peterseim. Variational multiscale stabilization and the exponential decay of fine-scale correctors. In *Building bridges: connections and challenges in modern approaches to numerical partial differential equations*, volume 114 of *Lect. Notes Comput. Sci. Eng.*, pages 341–367. Springer, Cham, 2016.

[36] D. Peterseim, D. Varga, and B. Verfürth. From domain decomposition to homogenization theory. In *DD25 proceedings (to appear)*. Springer, 2019.

[37] L. A. Richards. Capillary conduction of liquids through porous mediums. *Physics*, 1(5):318–333, 1931.

[38] M. van Genuchten. A closed form equations for predicting the hydraulic conductivity of unsaturated soils. *Soil Sci. Soc. Am. J.*, 40:892–898, 1980.

[39] J. Xu. Two-grid discretization techniques for linear and nonlinear PDEs. *SIAM J. Numer. Anal.*, 33(5):1759–1777, 1996.

[40] E. Zeidler. *Nonlinear functional analysis and its applications. IV. Applications to mathematical physics.* Springer-Verlag, New York, 1988.

A. L^2-error estimate for the Galerkin method

In this appendix, we prove an L^2-estimate for the Galerkin method (3.4).

Theorem A.1. *Let Assumptions 2.1 and 3.1 be fulfilled and suppose further that $A(x, \nabla u) = A_L(x, \nabla u, \nabla u)$. Let u be the solution to (2.2) and $u_{H,m}$ the solution to (3.7). Then it holds that*

$$\|u - u_{H,m}\|_0 \lesssim \left(H + C_{\mathrm{ol},m}^{1/2}\beta^m + \sup_{v_H \in V_H} |(\mathcal{Q}_m - \mathcal{Q}_m^u)v_H|_1\right)|u - u_{H,m}|_1 + \|\eta_{\mathrm{lin},1}(u, \widetilde{u}_{H,m})\|_{H^{-1}(\Omega)}$$

with the linearization error $\eta_{\text{lin},1}$ defined as

$$\langle \eta_{\text{lin},1}(u,v), \psi \rangle := \int_\Omega \big(A_L(x, \nabla u, \nabla v) - A(x, \nabla v) \big) \cdot \nabla \psi \, dx$$

and where $\widetilde{u}_{H,m}$ is defined via (A.1) below.

The first term of the above error estimate presumably is of $O(H^2)$ for the choice $m \approx |\log H|$ (up to the linearization error $|(\mathcal{Q}_m - \mathcal{Q}_m^u)v_H|_1$ discussed in Section 4.2). The linearization error $\eta_{\text{lin},1}(u,v)$ can be estimated in the same way as in Lemmas 4.4 and 4.5 by replacing u^* with u. According to those lemmas, we then need to estimate $|u - \widetilde{u}_{H,m}|_1$ and $\|\nabla(u - \widetilde{u}_{H,m})\|_{L^\infty(\Omega)}$. As shown below, $|u - \widetilde{u}_{H,m}|_1$ is of the same order as $|u - u_{H,m}|_1$ discussed in Theorem 4.2. However, it is not clear whether $W^{1,\infty}$-estimates for the multiscale space $V_{H,m}$ can be established in the spirit of [8, Chapter 8] to obtain a quadratic rate for $\eta_{\text{lin},1}$ in case of a Newton-type linearization.

Proof. Let $\widetilde{u}_{H,m} \in V_{H,m}$ be the unique solution to

$$\big(\mathfrak{A}^u \nabla \widetilde{u}_{H,m}, \nabla v_{H,m} \big)_\Omega = \big(\mathfrak{A}^u \nabla u, \nabla v_{H,m} \big)_\Omega \qquad \text{for all} \quad v_{H,m} \in V_{H,m}, \tag{A.1}$$

cf. [2]. Note that this can equivalently be written as

$$\big(A_L(x, \nabla u, \nabla \widetilde{u}_{H,m}), \nabla v_{H,m} \big)_\Omega = \big(A_L(x, \nabla u, \nabla u), \nabla v_{H,m} \big)_\Omega \qquad \text{for all} \quad v_{H,m} \in V_{H,m}.$$

We split the error $u - u_{H,m}$ into $u - \widetilde{u}_{H,m}$ and $\widetilde{u}_{H,m} - u_{H,m}$ and estimate both parts separately.

First step: Estimate of $\widetilde{u}_{H,m} - u_{H,m}$: By the monotonicity of A, Galerkin orthogonality, the assumption $A(x, \nabla u) = A_L(x, \nabla u, \nabla u)$ and (A.1), we deduce

$$
\begin{aligned}
|u_{H,m} - \widetilde{u}_{H,m}|_1^2 &\lesssim \big(A(x, \nabla u_{H,m}) - A(x, \nabla \widetilde{u}_{H,m}), \nabla(u_{H,m} - \widetilde{u}_{H,m}) \big)_\Omega \\
&= \big(A(x, \nabla u) - A(x, \nabla \widetilde{u}_{H,m}), \nabla(u_{H,m} - \widetilde{u}_{H,m}) \big)_\Omega \\
&= \big(A_L(x, \nabla u, \nabla u) - A(x, \nabla \widetilde{u}_{H,m}), \nabla(u_{H,m} - \widetilde{u}_{H,m}) \big)_\Omega \\
&= \big(A_L(x, \nabla u, \nabla \widetilde{u}_{H,m}) - A(x, \nabla \widetilde{u}_{H,m}), \nabla(u_{H,m} - \widetilde{u}_{H,m}) \big)_\Omega \\
&\leq \|\eta_{\text{lin},1}(u, \widetilde{u}_{H,m})\|_{H^{-1}(\Omega)} |u_{H,m} - \widetilde{u}_{H,m}|_1.
\end{aligned}
$$

The L^2-estimate is obtained by applying Friedrich's inequality.

Second step: Estimate of $u - \widetilde{u}_{H,m}$: This is in principle an L^2-estimate for a Galerkin LOD for an elliptic diffusion problem. The main issue, however, is that the multiscale space $V_{H,m}$ is not built with respect to the diffusion tensor $\mathfrak{A}^u = \mathbf{A}_L(x, \nabla u)$ but with respect to $\mathfrak{A} = \mathbf{A}_L(x, \nabla u^*)$. We first of all note that due to the projection, we have

$$|u - \widetilde{u}_{H,m}|_1 \lesssim \inf_{v_{H,m} \in V_{H,m}} |u - v_{H,m}|_1 \leq |u - u_{H,m}|_1.$$

Let now $z \in H_0^1(\Omega)$ and $z_{H,m} \in V_{H,m}$ be the solutions of the following dual problems

$$\big(\mathfrak{A}^u \nabla v, \nabla z \big)_\Omega = (v, u - \widetilde{u}_{H,m})_\Omega \qquad \text{for all} \quad v \in H_0^1(\Omega),$$

$$\text{and} \quad \big(\mathfrak{A}^u \nabla v_{H,m}, \nabla z_{H,m} \big)_\Omega = (v_{H,m}, u - \widetilde{u}_{H,m})_\Omega \qquad \text{for all} \quad v_{H,m} \in V_{H,m}.$$

Assumption 3.1 and Galerkin orthogonality imply

$$
\begin{aligned}
|z - z_{H,m}|_1 &\lesssim \inf_{v_{H,m} \in V_{H,m}} |z - v_{H,m}|_1 \leq |z - (\text{id} - \mathcal{Q}_m)I_H z|_1 \\
&\leq |z - (\text{id} - \mathcal{Q}^u)I_H z|_1 + |(\mathcal{Q}^u - \mathcal{Q}_m^u)I_H z|_1 + |(\mathcal{Q}_m^u - \mathcal{Q}_m)I_H z|_1,
\end{aligned}
$$

24

where the second term is estimated with Proposition 4.1. Employing (2.5) and the a priori (stability) estimate for z, the last term yields

$$|(\mathcal{Q}_m^u - \mathcal{Q}_m)I_H z|_1 \lesssim \left(\sup_{v_H \in V_H} |(\mathcal{Q}_m^u - \mathcal{Q}_m)v_H|_1\right)\|u - \widetilde{u}_{H,m}\|_0.$$

For the first term, we obtain with the ellipticity of $\mathfrak{A}^u$ as well as the definitions of $\mathcal{Q}^u$ and z that

$$
\begin{aligned}
|z - (\mathrm{id} - \mathcal{Q}^u)I_H z|_1^2 &\lesssim \left(\mathfrak{A}^u \nabla(z - (\mathrm{id} - \mathcal{Q}^u)I_H z), \nabla(z - (\mathrm{id} - \mathcal{Q}^u)I_H z)\right)_\Omega \\
&= \left(\mathfrak{A}^u \nabla z, \nabla(z - (\mathrm{id} - \mathcal{Q}^u)I_H z)\right)_\Omega \\
&= (u - \widetilde{u}_{H,m}, z - (\mathrm{id} - \mathcal{Q}^u)I_H z)_\Omega \\
&\lesssim H \|u - \widetilde{u}_{H,m}\|_0 \, |z - (\mathrm{id} - \mathcal{Q}^u)I_H z|_1.
\end{aligned}
$$

Combining the foregoing estimates, we conclude

$$|z - z_{H,m}|_1 \lesssim \left(H + C_{\mathrm{ol},m}^{1/2}\beta^m + \sup_{v_H \in V_H} |(\mathcal{Q}_m - \mathcal{Q}_m^u)v_H|_1\right)\|u - \widetilde{u}_{H,m}\|_0.$$

Finally, the definition of the dual problems yields

$$
\begin{aligned}
\|u - \widetilde{u}_{H,m}\|_0^2 &= \left(\mathfrak{A}^u \nabla(u - \widetilde{u}_{H,m}), \nabla z\right)_\Omega \\
&= \left(\mathfrak{A}^u \nabla(u - \widetilde{u}_{H,m}), \nabla(z - z_{H,m})\right)_\Omega \\
&\lesssim |z - z_{H,m}|_1 |u - \widetilde{u}_{H,m}|_1,
\end{aligned}
$$

which in combination with the already derived estimates finishes the proof. $\qquad\square$

B. Proof of Proposition 4.6

The proof of Proposition 4.6 simply relies on the definition of the element correctors and is similar to the results in [19, 20].

Proof of Proposition 4.6. Let $T \in \mathcal{T}_H$ and $v_H \in V_H$ be fixed. Abbreviate $w := (\mathcal{Q}_{T,m} - \mathcal{Q}_{T,m}^u)v_H$ and note that $w \in W(\mathrm{N}^m(T))$. We deduce by the definition of $\mathcal{Q}_{T,m}$ and $\mathcal{Q}_{T,m}^u$ that

$$
\begin{aligned}
|w|_1^2 &\lesssim \left(\widehat{\mathfrak{A}^u} \nabla w, \nabla w\right)_{\mathrm{N}^m(T)} \\
&= \left(\widehat{\mathfrak{A}^u} \nabla \mathcal{Q}_{T,m} v_H, \nabla w\right)_{\mathrm{N}^m(T)} - \left(\widehat{\mathfrak{A}} \nabla \mathcal{Q}_{T,m} v_H, \nabla w\right)_{\mathrm{N}^m(T)} + \left(\widehat{\mathfrak{A}} \nabla v_H, \nabla w\right)_T - \left(\widehat{\mathfrak{A}^u} \nabla v_H, \nabla w\right)_T \\
&= \left((\widehat{\mathfrak{A}} - \widehat{\mathfrak{A}^u})(\chi_T \nabla - \nabla \mathcal{Q}_{T,m})v_H, \nabla w\right)_{\mathrm{N}^m(T)} \\
&\leq \|(\widehat{\mathfrak{A}} - \widehat{\mathfrak{A}^u})(\chi_T \nabla - \nabla \mathcal{Q}_{T,m})v_H\|_{0,\mathrm{N}^m(T)} \, |w|_1.
\end{aligned}
$$

We then proceed as follows

$$
\begin{aligned}
|w|_1^2 &\lesssim \|(\widehat{\mathfrak{A}} - \widehat{\mathfrak{A}^u})(\chi_T \nabla - \nabla \mathcal{Q}_{T,m})v_H\|_{0,\mathrm{N}^m(T)}^2 \\
&\leq \max_{\psi|_T, \psi \in V_H} \frac{\|(\widehat{\mathfrak{A}} - \widehat{\mathfrak{A}^u})(\chi_T \nabla - \nabla \mathcal{Q}_{T,m})\psi\|_{0,\mathrm{N}^m(T)}^2}{|\psi|_{1,T}^2} |w|_{1,T}^2 \\
&\leq \sum_{T' \in \mathcal{T}_H, T' \subset \mathrm{N}^m(T)} \|\widehat{\mathfrak{A}} - \widehat{\mathfrak{A}^u}\|_{L^\infty(T')}^2 \max_{\psi|_T, \psi \in V_H} \frac{\|(\chi_T \nabla - \nabla \mathcal{Q}_{T,m})\psi\|_{0,T'}^2}{|\psi|_{1,T}^2} |w|_{1,T}^2 \\
&= E_{\mathcal{Q},T}^2 \, |w|_{1,T}^2,
\end{aligned}
$$

which finishes the proof. Note that only in the very first step we hide the lower spectral bound of $\widehat{\mathfrak{A}^u}|_{\mathrm{N}^m(T)}$ in the notation $\lesssim$, all other estimates are constant-free. $\qquad\square$