ESAIM: M2AN 48 (2014) 1331–1349
DOI: 10.1051/m2an/2013141

ESAIM: Mathematical Modelling and Numerical Analysis
www.esaim-m2an.org

A LOCALIZED ORTHOGONAL DECOMPOSITION METHOD FOR SEMI-LINEAR ELLIPTIC PROBLEMS *, **

Patrick Henning[1], Axel Målqvist[1] and Daniel Peterseim[2]

Abstract. In this paper we propose and analyze a localized orthogonal decomposition (LOD) method for solving semi-linear elliptic problems with heterogeneous and highly variable coefficient functions. This Galerkin-type method is based on a generalized finite element basis that spans a low dimensional multiscale space. The basis is assembled by performing localized linear fine-scale computations on small patches that have a diameter of order $H|\log(H)|$ where H is the coarse mesh size. Without any assumptions on the type of the oscillations in the coefficients, we give a rigorous proof for a linear convergence of the H^1-error with respect to the coarse mesh size even for rough coefficients. To solve the corresponding system of algebraic equations, we propose an algorithm that is based on a damped Newton scheme in the multiscale space.

Mathematics Subject Classification. 35J15, 65N12, 65N30.

Received November 14, 2012. Revised June 11, 2013.
Published online August 13, 2014.

1. Introduction

This paper is devoted to the numerical approximation of solutions of semi-linear elliptic problems with rapidly oscillating and highly varying coefficient functions. We are concerned with second-order partial differential equations of the type

$$-\nabla \cdot (A\nabla u) + F(u, \nabla u) = g$$

with prescribed (zero-) Dirichlet boundary condition for the unknown function u. Here, g is a given source term, A is a given highly variable diffusion matrix and F is a given highly variable nonlinear term that represents advective and reactive processes. In particular, we have a linear term of second order and nonlinear terms of order 1 and 0. A typical application is the stationary (Kirchhoff transformed) Richards equation that describes the groundwater flow in unsaturated soils (*cf.* [1,4,5]). The corresponding equation for the unknown generalized

Keywords and phrases. Finite element method, *a priori* error estimate, convergence, multiscale method, non-linear, computational homogenization, upscaling.

* *A. Målqvist and P. Henning are supported by The Göran Gustafsson Foundation and The Swedish Research Council.*

** *The research of D. Peterseim was supported by the Humboldt-Universität and the DFG Research Center Matheon Berlin.*

[1] Department of Information Technology, Uppsala University, Box 337, 75105 Uppsala, Sweden.

[2] Institut für Numerische Simulation der Universität Bonn, Wegelerstr. 66, 53123 Bonn, Germany.
patrick.henning@uni-muenster.de

Article published by EDP Sciences

pressure u reads

$$\nabla \cdot (K \nabla u) - \nabla \cdot (K\, kr(M(u))\vec{e}) = g,$$

where K is the hydraulic conductivity in the soil, kr the relative permeability depending on the saturation, M is some nonlinearity arising from the Kirchhoff transformation and $\vec{e}$ denotes the gravity vector. If we add an infiltration process, the equation receives an additional nonlinear reaction term.

The numerical treatment of such equations is often complicated and expensive. Due to the high variability of the coefficient functions, one requires extremely fine computational grids that are able to capture all the fine scale oscillations. Using standard methods such as Finite Element or Finite Volume schemes, this results in systems of equations of enormous size and therefore in a tremendous computational demand that can not be handled in a lot of scenarios.

Multiscale methods aim to overcome this difficulty by decoupling the fine scale computations into local parts. Prominent examples of multiscale methods are the Heterogeneous Multiscale Method (HMM) by E and Engquist [13] and the Multiscale Finite Element Method (MsFEM) proposed by Hou and Wu [20]. Both methods fit into a common framework and are strongly related to numerical homogenization (*cf.* [14,15,18]). HMM and MsFEM are typically not constructed for a direct approximation of exact solutions but for homogenized solutions and corresponding correctors instead. This implies that they are only able to approximate the exact solution up to a modeling error that depends crucially on the homogenization setting (*cf.* [14]). In the absence of strong assumptions like periodicity and scale separation, accurate approximations are therefore hard to achieve.

We are concerned with a multiscale method that is based on the concept of the Variational Multiscale Method (VMM) proposed by Hughes *et al.* [21]. In comparison to HMM and MsFEM, the VMM aims to a direct approximation of the exact solution without suffering from a modeling error remainder arising from homogenization theory. The key idea of the Variational Multiscale Method is to construct a splitting of the original solution space V into the direct sum of a low dimensional space for coarse grid approximations and high dimensional space for fine scale reconstructions. In this work, we consider a modification and extension of this idea that was developed in [27,30] and that was explicitly proposed in [31]. Here, the splitting is such that we obtain an accurate but low dimensional space V^{ms} (where we are looking for our fine scale approximation instead of an approximation of a coarse part) and a high dimensional residual space V^{f}. The construction of V^{ms} involves the computation of one fine scale problem in a small patch per degree of freedom. Mesh-adaptive versions of the VMM with patch size control are discussed in [27–29,33]. The first rigorous proof of convergence was recently obtained in [31] for linear diffusion problems under minimal regularity assumptions.

In this contribution, we present an efficient way of handling semi-linear elliptic multiscale problems in the modified VMM framework, including a proof of convergence based on the techniques established in [31]. Even though the original problem is nonlinear, the local fine scale problems are purely linear that can be solved in parallel. The main result of this article is the optimal convergence of the H^1-error between exact solution u and its multiscale approximation u_H^{ms}. We show that, if the patch size is of order $H|\log(H)|$, the following error bound

$$\|u - u_H^{\mathrm{ms}}\|_{H^1(\Omega)} \leq CH$$

holds with a generic constant C independent of the mesh size of the computational grid and the oscillations of A and F.

The paper is structured as follows. In Section 2 we introduce the setting of this paper, including the assumptions on the considered semi-linear problem. In Section 3 we present and motivate our method and we state the corresponding optimal convergence result. This result is then proved in Section 4. In Section 5, we propose an algorithm for the solution of the arising nonlinear algebraic equations. This algorithm is based on a damped Newton scheme in the multiscale space. Finally, Section 6 supports the theoretical results by a numerical experiment.

2. Setting

Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain with polyhedral boundary, let $V := H_0^1(\Omega)$ and let $A \in L^\infty(\Omega, \mathbb{R}^{d\times d}_{\mathrm{sym}})$ denote a matrix valued function with uniformly strictly positive eigenvalues. We assume that the space $H_0^1(\Omega)$ is endowed with the H^1-semi norm given by $|v|_{H^1(\Omega)} := \|\nabla v\|_{L^2(\Omega)}$ (which is equivalent to the common H^1-norm in $H_0^1(\Omega)$). By $\langle\cdot,\cdot\rangle := (\cdot,\cdot)_{L^2(\Omega)}$ we denote the inner product in $L^2(\Omega)$ and $F : \Omega\times\mathbb{R}\times\mathbb{R}^d \to \mathbb{R}$ is a nonlinear measurable function.

Given some source term $g \in L^2(\Omega) \subset H^{-1}(\Omega)$ we are concerned to find $u \in H_0^1(\Omega)$ (*i.e.* with a homogeneous Dirichlet boundary condition) with

$$\langle A\nabla u, \nabla v\rangle + \langle F(\cdot, u, \nabla u), v\rangle = \langle g, v\rangle \tag{2.1}$$

for all test functions $v \in H_0^1(\Omega)$. To simplify the notation, we define the operator $B : H_0^1(\Omega) \to H^{-1}(\Omega)$ by

$$\langle B(v), w\rangle_{H^{-1},H_0^1} := \langle A\nabla v, \nabla w\rangle + \langle F(\cdot, v, \nabla v), w\rangle \quad \text{for } v, w \in H_0^1(\Omega),$$

where $\langle\cdot,\cdot\rangle_{H^{-1},H_0^1}$ denote the dual pairing in $H_0^1(\Omega)$.

Here, the diffusion diffusion matrix A may be strongly heterogeneous and highly variable. The non-linearity $F(\cdot,\xi,\zeta)$ may as well oscillate rapidly without any assumptions on the type of the oscillations. One application can be the Richards equation, which we will discuss more in Section 6.

However, we assume implicitly that the lower-order term F does not dominate the equation. In this regime, it is sufficient to construct a multiscale space independent of the non-linearity by solving local linear problems on the fine scale. If the lower-order term is dominant, some constants in our error analysis will be large and the proposed method needs modifications with respect to the construction of the multiscale basis. A typical example where the lower-order term is dominant is the modeling of transport of solutes in groundwater where one has to deal with extremely large Péclet numbers and a corresponding scaling of the advective terms. In this case, the resolution of oscillations of F is necessary for accurate upscaled and homogenized approximation (*cf.* [16, 17]).

For the subsequent analytical considerations and in order to guarantee a unique solution of (2.1), we make the following assumptions.

Assumption 1.

(A1) $A \in L^\infty\left(\Omega, \mathbb{R}^{d\times d}_{\mathrm{sym}}\right)$ with

$$\infty > \beta := \|A\|_{L^\infty(\Omega)} = \operatorname*{ess\,sup}_{x\in\Omega} \sup_{\zeta\in\mathbb{R}^d\setminus\{0\}} \frac{A(x)\zeta\cdot\zeta}{|\zeta|^2}.$$

and there exists α such that

$$0 < \alpha := \operatorname*{ess\,inf}_{x\in\Omega} \inf_{\zeta\in\mathbb{R}^d\setminus\{0\}} \frac{A(x)\zeta\cdot\zeta}{|\zeta|^2},$$

(A2) There exist $L_1, L_2 \in \mathbb{R}_{>0}$ such that uniformly for almost every x in Ω:

$$\begin{aligned} |F(x,\xi_1,\zeta) - F(x,\xi_2,\zeta)| &\le L_1|\xi_1-\xi_2|, \quad \text{for all } \zeta\in\mathbb{R}^d,\ \xi_1,\xi_2\in\mathbb{R},\\ |F(x,\xi,\zeta_1) - F(x,\xi,\zeta_2)| &\le L_2|\zeta_1-\zeta_2|, \quad \text{for all } \zeta_1,\zeta_2\in\mathbb{R}^d,\ \xi\in\mathbb{R},\\ F(x,0,0) &= 0. \end{aligned}$$

(A3) B is strongly monotone, *i.e.* there exist $c_0 > 0$ so that for all $u, v \in H_0^1(\Omega)$:

$$\langle B(u) - B(v), u - v\rangle_{H^{-1},H_0^1} \ge c_0|u-v|^2_{H^1(\Omega)}. \tag{2.2}$$

Under assumptions (A1)−(A3), the Browder−Minty theorem (*cf.* [36], Sect. 3, Thm. 1.5 therewithin) yields a unique solution of problem (2.1).

Typically, the validity of Assumption (A3) can be checked by looking at the properties of the nonlinear function F. For instance, if there exists a constant $\alpha_0 \geq 0$, such that $\partial_\xi F(x,\xi,\zeta) \geq \alpha_0$ for all ζ and almost every x (*i.e.* $F(x,\cdot,\zeta)$ is monotonically increasing) and if α_0 and L_2 are such that $L_2 \leq 2\alpha_0$ and $L_2 < 2\alpha$ then (A3) is fulfilled. This can be checked by a simple calculation:

$$\begin{aligned}\langle B(u) - B(v), u - v\rangle_{H^{-1},H_0^1} &\geq \alpha||\nabla u - \nabla v||^2_{L^2(\Omega)} + \alpha_0||u-v||^2_{L^2(\Omega)} - L_2(|u-v|,|\nabla u - \nabla v|)_{L^2(\Omega)} \\ &\geq \left(\alpha - \frac{L_2}{2}\right)||\nabla u - \nabla v||^2_{L^2(\Omega)} + \left(\alpha_0 - \frac{L_2}{2}\right)||u-v||^2_{L^2(\Omega)}.\end{aligned}$$

Remark 2.1. Let $C_\Omega < \operatorname{diam}\Omega$ denote the optimal constant in the Friedrichs inequality for $H_0^1(\Omega)$ functions. Observe that (A1)−(A3) imply that the solution $u \in H_0^1(\Omega)$ of (2.1) fulfills

$$\begin{aligned}&\|F(u,\nabla u)\|_{L^2(\Omega)} \leq \|F(u,\nabla u) - F(0,\nabla u)\|_{L^2(\Omega)} + \|F(0,\nabla u) - F(0,0)\|_{L^2(\Omega)} \\ &\leq (L_1 C_\Omega + L_2)|u|_{H^1(\Omega)} \leq C_\Omega \frac{L_1 C_\Omega + L_2}{c_0}\|g\|_{L^2(\Omega)}.\end{aligned} \tag{2.3}$$

Note that problem (2.1) also covers equations such as

$$-\nabla\cdot(\kappa(u)A\nabla u) + F(u,\nabla u) = g,$$

for a strictly positive and sufficiently regular function κ (independent of x). In this case, the equation can be rewritten as

$$-\nabla\cdot A\nabla u + \tilde{F}(u,\nabla u) = \tilde{g}.$$

In the remainder of this paper, we use the notation $q_1 \lesssim q_2$ if $q_1 \leq Cq_2$ where $C > 0$ is a constant that only depends on the shape regularity of the mesh, but not on the mesh size. Dependencies such as $(L_1 + L_2)\alpha^{-1}$ are always explicitly stated whereas dependencies on the contrast $\frac{\beta}{\alpha}$ are allowed to be contained in the notation $\lesssim$ for the sake of simplicity.

3. Multiscale method

In this section we propose a local orthogonal decomposition (LOD) method that is based on the concept introduced by Hughes *et al.* [21,22] and the specific constructions proposed in [27,30] for linear problems. The required multiscale (MS) basis functions are obtained with the strategy established in [31].

The main idea of the Variational Multiscale Method is to start from a finite element space $\mathcal{V}_h$ with a highly resolved computational grid and to construct a splitting of this space into the direct sum $\mathcal{V}_h = \mathcal{V}^{\mathrm{l}} \oplus \mathcal{V}^{\mathrm{f}}$ of a low dimensional space $\mathcal{V}^{\mathrm{l}}$ and a "detail space" $\mathcal{V}^{\mathrm{f}}$ containing all the missing oscillations. Then, a basis of $\mathcal{V}^{\mathrm{l}}$ is assembled and we can compute a Galerkin approximation u_{l} of u in $\mathcal{V}^{\mathrm{l}}$. However, the success of this approach strongly depends on the choice of $\mathcal{V}^{\mathrm{l}}$. On the one hand, the costs for assembling a basis of $\mathcal{V}^{\mathrm{l}}$ must be kept low. On the other hand, the basis functions somehow need to contain information about fine scale features. For instance, a standard coarse finite element space is cheap to assemble but will fail to yield reliable approximations. On the contrary, the space spanned by high resolution finite element approximations yields perfect approximations, but is as costly as the original problem that we tried to avoid. Therefore, the key is to find an optimal balance between costs and accuracy. In previous works (*cf.* [21,27,28]) the multiscale basis (MS-basis) of $\mathcal{V}^{\mathrm{l}}$ was constructed involving the full multiscale operator B that corresponds with the left hand side of the original problem. In a fully linear setting, this can be a reasonable choice. However, it gets extremely expensive if B is a nonlinear operator, since it leads to numerous nonlinear equations to solve. Furthermore it is not clear if the constructed set of basis functions leads to good approximations. One novelty of this work is

that we do not involve the full operator B in the construction of the MS-basis, but only the linear diffusive part $\langle A\nabla\cdot, \nabla\cdot\rangle$. Even though the oscillations of F are not captured by the MS-basis, we can show that we are still able to obtain accurate approximations and to preserve the optimal convergence rates.

3.1. Notation and discretization

Let $\mathcal{T}_H$ denote a regular triangulation of Ω and let $H : \overline{\Omega} \to \mathbb{R}_{>0}$ denote the $\mathcal{T}_H$-piecewise constant mesh size function with $H|_T = H_T := \operatorname{diam}(T)$ for all $T \in \mathcal{T}_H$. Additionally, let $\mathcal{T}_h$ be a regular triangulation of Ω that is supposed to be a refinement of $\mathcal{T}_H$. We assume that $\mathcal{T}_h$ is sufficiently small so that all fine scale features of B are captured by the mesh. The mesh size h denotes the maximum diameter of an element of $\mathcal{T}_h$. The corresponding classical (conforming) finite element spaces of continuous piecewise polynomials of degree 1 are given by

$$V_H := \left\{v_H \in H_0^1(\Omega) \mid \forall T \in \mathcal{T}_H : \ (v_H)|_T \text{ is affine}\right\},$$
$$V_h := \left\{v_h \in H_0^1(\Omega) \mid \forall K \in \mathcal{T}_h : \ (v_h)|_K \text{ is affine}\right\}.$$

By J, we denote the dimension of V_H and by $\mathcal{N}_H = \{z_j | \, 1 \le j \le J\}$ the set of interior vertices of $\mathcal{T}_H$. For every vertex $z_j \in \mathcal{N}_H$, let $\lambda_j \in V_H$ denote the associated nodal basis function (tent function), *i.e.* $\lambda_j \in V_H$ with the property $\lambda_j(z_i) = \delta_{ij}$ for all $1 \le i, j \le J$.

From now on, we denote by $u_h \in V_h$ the classical finite element approximation of u in the discrete (highly resolved) space V_h, *i.e.* $u_h \in V_h$ solves

$$\int_\Omega A\nabla u_h \cdot \nabla v_h + F(\cdot, u_h, \nabla u_h)v_h = \int_\Omega g v_h \tag{3.1}$$

for all $v_h \in V_h$. We assume that V_h resolves the micro structure such that the error $\|u - u_h\|_{H^1(\Omega)}$ falls below a given tolerance. For standard finite element methods the error typically scales like $C \cdot h^s$ for some $s \ge \frac{1}{2}$. However, for regular coefficients, C depends on the derivative of A with respect to the spatial variable. If A oscillates rapidly, the derivatives become very large and h must be very small to compensate the dominance of C. This is only fulfilled, when h resolves the micro structure (we refer to [34,35] for some quantitative characterization of this so-called resolution condition). We are therefore dealing with pre-asymptotic effects for the standard methods. The multiscale method that we propose in the subsequent sections is designed to approximate u_h with an error proportional to the coarse mesh size H independent of fine scale oscillations of the data or the regularity of the solution, *i.e.*, we do not have such pre-asymptotic effects.

3.2. Quasi interpolation

The key tool in our construction is a linear (quasi-)interpolation operator $\mathfrak{I}_H : V_h \to V_H$ that is continuous and surjective. The kernel of this operator is going to be our fine space (or remainder space) V_h^{f}. In [31] a weighted Clément interpolation operator was used. In this work, we do not specify the choice. Instead, we state a set of assumptions that must be fulfilled in order to derive an optimal approximation result for the constructed multiscale method.

Assumption 2. (Assumptions on the quasi-interpolation operator)**.**

(A4) $\mathfrak{I}_H \in L(V_h, V_H)$, *i.e.* $\mathfrak{I}_H$ is linear,

(A5) the restriction of $\mathfrak{I}_H$ to V_H is an isomorphism with L^2-stable inverse $(\mathfrak{I}_H|_{V_H})^{-1}$, *i.e.* $\|(\mathfrak{I}_H|_{V_H})^{-1}(v_H)\|_{L^2(\Omega)} \le C_{\mathfrak{I}_H^{-1}}\|v_H\|_{L^2(\Omega)}$ for all $v_H \in V_H$ and with a generic constant $C_{\mathfrak{I}_H^{-1}}$ only depending on the shape regularity of $\mathcal{T}_H$ and $\mathcal{T}_h$.

(A6) there exists a generic constant $C_{\mathfrak{I}_H}$, only depending on the shape regularity of $\mathcal{T}_H$ and $\mathcal{T}_h$, such that for all $v_h \in V_h$ and for all $T \in \mathcal{T}_H$ there holds

$$H_T^{-1}\|v_h - \mathfrak{I}_H(v_h)\|_{L^2(T)} + \|\nabla(v_h - \mathfrak{I}_H(v_h))\|_{L^2(T)} \le C_{\mathfrak{I}_H}\|\nabla v_h\|_{L^2(\omega_T)}$$

with

$$\omega_T := \bigcup\{K \in \mathcal{T}_H \,|\, \overline{K} \cap \overline{T} \neq \emptyset\}.$$

(A7) there exists a generic constant $C'_{\mathfrak{I}_H}$, only depending on the shape regularity of $\mathcal{T}_H$ and $\mathcal{T}_h$, such that for all $v_H \in V_H$ there exists $v_h \in V_h$ with

$$\mathfrak{I}_H(v_h) = v_H, \quad |v_h|_{H^1(\Omega)} \leq C'_{\mathfrak{I}_H}|v_H|_{H^1(\Omega)} \quad \text{and} \quad \operatorname{supp} v_h \subset \operatorname{supp} v_H.$$

Observe that (A6) limits the growth of the support of an $v_H \in V_H$ when $\mathfrak{I}_H$ is applied to it, *i.e.* $\operatorname{supp}(I_H(v_H)) = \bigcup\{K \in \mathcal{T}_H \,|\, \overline{K} \cap \operatorname{supp}(v_H) \neq \emptyset\}$. We also note that the classical nodal interpolation operator does not fulfill assumption (A6) for $d > 1$ because the constant $C_{\mathfrak{I}_H}$ blows up for $h \to 0$. Numerical experiments confirm that such a choice leads in fact to instabilities in the later method. One possibility is to choose $\mathfrak{I}_H$ as a weighted Clément interpolation operator. This construction was proposed in [31]. Given $v \in H^1_0(\Omega)$, $\mathfrak{I}_H v := \sum_{j=1}^J v_j \lambda_j$ defines a (weighted) Clément interpolant with nodal values

$$v_j := \left(\int_\Omega v\lambda_j \,\mathrm{d}x\right) / \left(\int_\Omega \lambda_j \,\mathrm{d}x\right) \tag{3.2}$$

for $1 \leq j \leq J$ (*cf.* [11]) and zero in the boundary nodes. Furthermore, there exists the desired generic constant $C_{\mathfrak{I}_H}$ (only depending on the mesh regularity parameter and in particular independent of H_T) such that for all $v \in H^1_0(\Omega)$ and for all $T \in \mathcal{T}_H$ there holds

$$H_T^{-1}\|v - \mathfrak{I}_H v\|_{L^2(T)} + \|\nabla(v - \mathfrak{I}_H v)\|_{L^2(T)} \leq C_{\mathfrak{I}_H}\|\nabla v\|_{L^2(\omega_T)}.$$

We refer to [11] for a proof of this estimate. This gives us (A6). Assumption (A4) is obvious. The validity of (A5) and (A7) was proved in [31].

Note that in certain applications, additional features (*e.g.*, orthogonality properties) of the chosen interpolation operator may be exploited for improved error estimates (see, *e.g.*, [31] Rem. 3.2 and [10]).

3.3. Multiscale splitting and modified nodal basis

In this section, we construct a splitting of the high resolution finite element space V_h into a low dimension multiscale space V^{ms} and some high dimensional remainder space V_h^{f}. From now on, we let $\mathfrak{I}_H : V_h \to V_H$ denote an interpolation operator fulfilling the properties (A4)–(A7). Recall that $V_H \subset V_h$. We start with defining V_h^{f} as the kernel of $\mathfrak{I}_H$ in V_h:

$$V_h^{\mathrm{f}} := \{v_h \in V_h \mid \mathfrak{I}_H v_h = 0\}.$$

V_h^{f} represents the features in V_h not captured by V_H. Using assumption (A5) we get

$$V_h = V_H \oplus V_h^{\mathrm{f}}, \quad \text{where } \underbrace{v_h}_{\in V_h} = \underbrace{(\mathfrak{I}_H|_{V_H})^{-1}(\mathfrak{I}_H(v_h))}_{\in V_H} + \underbrace{v_h - (\mathfrak{I}_H|_{V_H})^{-1}(\mathfrak{I}_H(v_h))}_{\in V_h^{\mathrm{f}}}. \tag{3.3}$$

Here, the property $(\mathfrak{I}_H \circ (\mathfrak{I}_H|_{V_H})^{-1}))(v_H) = v_H$ for all $v_H \in V_H$ implies the equation $\mathfrak{I}_H(v_h - (\mathfrak{I}_H|_{V_H})^{-1}(\mathfrak{I}_H(v_h))) = \mathfrak{I}_H(v_h) - (\mathfrak{I}_H \circ (\mathfrak{I}_H|_{V_H})^{-1})(\mathfrak{I}_H(v_h)) = 0$. We still need to modify the splitting of V_h, because V_H is an inappropriate space for a multiscale approximation. We therefore look for the orthogonal complement of V_h^{f} in V_h with respect to the inner product $\langle A\nabla\cdot, \nabla\cdot\rangle_{L^2(\Omega)}$. For this purpose, we define the orthogonal projection $P^{\mathrm{f}} : V_h \to V_h^{\mathrm{f}}$ as follows. For a given $v_h \in V_h$, $P^{\mathrm{f}}(v_h) \in V_h^{\mathrm{f}}$ solves

$$\left\langle A\nabla P^{\mathrm{f}}(v_h), \nabla w^{\mathrm{f}}\right\rangle = \left\langle A\nabla v_h, \nabla w^{\mathrm{f}}\right\rangle \quad \text{for all } w^{\mathrm{f}} \in V_h^{\mathrm{f}}.$$

Defining the multiscale space $V_{H,h}^{\mathrm{ms}}$ by $V_{H,h}^{\mathrm{ms}} := (1 - P^{\mathrm{f}})(V_H)$, this directly leads to the orthogonal decomposition

$$V_h = V_{H,h}^{\mathrm{ms}} \oplus V_h^{\mathrm{f}}, \tag{3.4}$$

because

$$V_h = \operatorname{kern}(P^{\mathrm{f}}) \oplus V_h^{\mathrm{f}} = (1-P^{\mathrm{f}})(V_h) \oplus V_h^{\mathrm{f}} \overset{(3.3)}{=} (1-P^{\mathrm{f}})(V_H) \oplus V_h^{\mathrm{f}} = V_{H,h}^{\mathrm{ms}} \oplus V_h^{\mathrm{f}}.$$

Hence, any function $v_h \in V_h$ can be decomposed into $v_h = v_H^{\mathrm{ms}} + v^{\mathrm{f}}$ with $v_H^{\mathrm{ms}} = (\mathfrak{I}_H|_{V_H})^{-1}(\mathfrak{I}_H(v_h)) - P^{\mathrm{f}}((\mathfrak{I}_H|_{V_H})^{-1}(\mathfrak{I}_H(v_h)))$ and $v^{\mathrm{f}} = v_h - (\mathfrak{I}_H|_{V_H})^{-1}(\mathfrak{I}_H(v_h)) + P^{\mathrm{f}}((\mathfrak{I}_H|_{V_H})^{-1}(\mathfrak{I}_H(v_h)))$. Furthermore it holds $\langle A\nabla v_H^{\mathrm{ms}}, \nabla w^{\mathrm{f}}\rangle = 0$ for all $w^{\mathrm{f}} \in V_h^{\mathrm{f}}$. The space $V_{H,h}^{\mathrm{ms}}$ is a multiscale space of the same dimension as the coarse space V_H. However, note that it is only constructed on the basis of the oscillations of A. The oscillations of F are not taken into account. We will show that $V_{H,h}^{\mathrm{ms}}$ still yields the desired approximation properties.

We now introduce a basis of $V_{H,h}^{\mathrm{ms}}$. The image of the nodal basis function $\lambda_j \in V_H$ under the fine scale projection P^{f} is denoted by $\phi_j^h = P^{\mathrm{f}}(\lambda_j) \in V_h^{\mathrm{f}}$, *i.e.*, ϕ_j^h satisfies the corrector problem

$$\langle A\nabla\phi_j^h, \nabla w\rangle = \langle A\nabla\lambda_j, \nabla w\rangle \quad \text{for all } w \in V_h^{\mathrm{f}}. \tag{3.5}$$

A basis of $V_{H,h}^{\mathrm{ms}}$ is then given by the modified nodal basis

$$\left\{\lambda_j^{\mathrm{ms}} := \lambda_j - \phi_j^h \mid 1 \le j \le J\right\}. \tag{3.6}$$

As we can see, solving (3.5) involves a fine scale computation on the whole domain Ω. However, since the right hand side has small support, we are able to localize the computations. As we will see in the next section, the correctors show an exponential decay outside of the support of the coarse shape function λ_j.

First, we define a multiscale approximation that is based on the above orthogonal decomposition of V_h, but without localization.

Definition 3.1 (Multiscale approximation without localization)**.** The Galerkin approximation $u_{H,h}^{\mathrm{ms}} \in V_{H,h}^{\mathrm{ms}}$ of the exact solution u of problem (2.1) is defined as the solution of

$$\langle A\nabla u_{H,h}^{\mathrm{ms}}, \nabla v\rangle + \langle F\left(u_{H,h}^{\mathrm{ms}}, \nabla u_{H,h}^{\mathrm{ms}}\right), v\rangle = \langle g, v\rangle \quad \text{for all } v \in V_{H,h}^{\mathrm{ms}}. \tag{3.7}$$

3.4. Localization

So far, in order to construct a suitable multiscale space, we derived a set of linear fine scale problems (3.5) that can be solved in parallel. Still, as already mentioned in the previous section, these corrector problems are fine scale equations formulated on the whole domain Ω which makes them almost as expensive as the original problem. However, in [31] it was shown that the correction ϕ_j^h decays exponentially outside of the support of the coarse basis function λ_j. We specify this feature as follows. Let $k \in \mathbb{N}_{>0}$. We define nodal patches $\omega_{j,k}$ of k coarse grid layers centered around the node $z_j \in \mathcal{N}_H$ by

$$\begin{aligned} \omega_{j,1} &:= \operatorname{supp}\lambda_j = \cup\left\{T \in \mathcal{T}_H \mid z_j \in \overline{T}\right\}, \\ \omega_{j,k} &:= \cup\left\{T \in \mathcal{T}_H \mid \overline{T} \cap \overline{\omega}_{j,k-1} \neq \emptyset\right\} \quad \text{for } k \ge 2. \end{aligned} \tag{3.8}$$

These are the truncated computational domains for the corrector problems (3.5). The fast decay is summarized by the following lemma.

Lemma 3.2 (Decay of the local correctors [31])**.** *Let assumptions* (A1) *and* (A4)−(A7) *be fulfilled. Then, for all nodes $z_j \in \mathcal{N}_H$ and for all $k \in \mathbb{N}_{>0}$, the correctors ϕ_j^h satisfy the estimates*

$$\|A^{1/2}\nabla\phi_j^h\|_{L^2(\Omega\setminus\omega_{j,k})} \lesssim \mathrm{e}^{-rk}\|A^{1/2}\nabla\phi_j^h\|_{L^2(\Omega)}$$

with a generic rate r that is proportional to $(\alpha/\beta)^{1/2}$ but independent of variations of A. Recall the definition of '$\lesssim$' at the end of Section 2.

This fast decay motivates an approximation of ϕ_j^h on the truncated nodal patches $\omega_{j,k}$. We therefore define localized fine scale spaces by intersecting V_h^{f} with those functions that vanish outside the patch $\omega_{j,k}$, *i.e.*

$$V_h^{\mathrm{f}}(\omega_{j,k}) := \left\{ v \in V_h^{\mathrm{f}} \mid v_{|\Omega\setminus\omega_{j,k}} = 0 \right\}$$

for a given node $z_j \in \mathcal{N}_H$. The solutions $\phi_{j,k}^h \in V_h^{\mathrm{f}}(\omega_{j,k})$ of

$$\langle A\nabla\phi_{j,k}^h, \nabla w\rangle = \langle A\nabla\lambda_j, \nabla w\rangle \quad \text{for all } w \in V_h^{\mathrm{f}}(\omega_{j,k}), \tag{3.9}$$

are approximations of ϕ_j^h from (3.5) with local support and therefore cheap to solve. We define localized multiscale finite element spaces by

$$V_{H,h}^{\mathrm{ms},k} = \operatorname{span}\left\{ \lambda_{j,k}^{\mathrm{ms}} := \lambda_j - \phi_{j,k}^h \mid 1 \le j \le J \right\} \subset V_h. \tag{3.10}$$

We can now define a LOD approximation by localizing the corrector problems for the basis functions.

Definition 3.3 (LOD approximation)**.** The Galerkin approximation $u_{H,h}^{\mathrm{ms},k} \in V_{H,h}^{\mathrm{ms},k}$ of the exact solution u of problem (2.1) is defined as the solution of

$$\left\langle A\nabla u_{H,h}^{\mathrm{ms},k}, \nabla v \right\rangle + \left\langle F\left(u_{H,h}^{\mathrm{ms},k}, \nabla u_{H,h}^{\mathrm{ms},k}\right), v \right\rangle = \langle g, v\rangle \quad \text{for all } v \in V_{H,h}^{\mathrm{ms},k}. \tag{3.11}$$

Note, that changing the data functions F and g does not change the multiscale basis $\{\lambda_{j,k}^{\mathrm{ms}} \mid 1 \le j \le J\}$. Once $V_{H,h}^{\mathrm{ms},k}$ is computed, it can be reused for various combinations of F and g. This makes the new problems cheap to solve.

Remark 3.4. Observe that we never need to solve a problem on the scale of the oscillations of $F(\cdot,\xi,\zeta)$ in the case that they are faster than the oscillations of $A(\cdot)$. However, we implicitly assume that the arising integrals can be computed exactly (or with high accuracy). Practically this implies that a sufficiently high quadrature rule must be used. So even if the fine grid is not fine enough to resolve the variations of F, at least the quadrature rule must be fine enough to capture the correct averaged values. From Theorem 3.5 below we deduce that the influence of the oscillations of $F(\cdot,\xi,\zeta)$ remains small, as long as we have an accurate approximation of the averages on each coarse grid element. A similar observation holds for standard finite elements, where classical convergence rates can be expected as soon as the oscillations of A are resolved by the fine grid (independent of the oscillations of F).

3.5. *A priori* error estimate

We are now prepared to state the main result of this article, namely the optimal convergence of the method for the case that the local patches $\omega_{j,k}$ have a diameter of order $H|\log(H)|$.

Theorem 3.5. *Let $u \in H_0^1(\Omega)$ denote the exact solution given by problem (2.1), let $u_h \in V_h$ denote the corresponding finite element approximation in the Lagrange space with a highly resolved computational grid (*i.e. *the solution of (3.1)) and let $u_{H,h}^{\mathrm{ms},k} \in V_{H,h}^{\mathrm{ms},k}$ be the solution of our proposed multiscale method with localization* (i.e. *the solution of (3.11)). If assumptions* (A1)–(A7) *are satisfied and if $k \gtrsim |\log(\|H\|_{L^\infty(\Omega)})|$, then the a priori error estimate*

$$\left\| u - u_{H,h}^{\mathrm{ms},k} \right\|_{H^1(\Omega)} \le C\left(L_1, L_2, \alpha, \beta, c_0\right)\left(\|H\|_{L^\infty(\Omega)} + \|u - u_h\|_{H^1(\Omega)}\right).$$

holds with a generic constant C that does not depend on mesh sizes and oscillations of A and F. A suitable choice of the localization parameter k depends on the square root of the contrast, i.e. *the multiplicative constant hidden in $k \approx |\log(\|H\|_{L^\infty(\Omega)})|$ is proportional to $\sqrt{\frac{\beta}{\alpha}}$.*

A proof of Theorem 3.5 is presented in the subsequent section. In particular, the result is a conclusion from Theorem 4.3 which is stated in Section 4 below. In Theorem 4.3 we also give details on the generic constant C. We will see that it essentially depends on $\frac{(L_1+L_2)}{\alpha}$. Recall that L_1 and L_2 denote the Lipschitz constants of F (*cf.* (A2)) and that α is the smallest eigenvalue of A. This shows the significance of assuming that the problem is not dominated by the lower order term. For instance, consider the scenario of a pollutant being transported by groundwater flow. In this case, A describes the hydraulic conductivity which changes its properties on a scale of size ϵ. On the other hand, F describes the gravity driven flow that is scaled with the so called Péclet number. However, in the described scenario the Péclet number is of order ϵ^{-1} (*cf.* Bourlioux and Majda [7]) implying that $\mathrm{O}(L_1) = \epsilon^{-1}$. So the generic constant C is of order ϵ^{-1}. This means that we need $H < \epsilon$, *i.e.* we still need to resolve the micro structure with the coarse grid $\mathcal{T}_H$ producing the same costs as the original problem. If $H \gg \epsilon$ the estimate stated in Theorem 3.5 is of no value, because the right hand side remains large.

4. Error analysis

This section is devoted to the proof of Theorem 3.5. In particular, we state a detailed version of the result (see Thm. 4.3 below), where we specify the occurring constants. The proof is splitted into several lemmata. We start with an *a priori* error estimate for the multiscale approximation without localization.

Lemma 4.1. *Let $u_h \in V_h$ denote the highly resolved finite element approximation defined* via *equation* (3.1) *and let $u_{H,h}^{\mathrm{ms}} \in V_{H,h}^{\mathrm{ms}}$ denote the LOD approximation given by equation* (3.7). *Under assumptions* (A1)–(A7), *the* a priori *error estimate*

$$|u_h - u_H^{\mathrm{ms}}|_{H^1(\Omega)} \lesssim \tilde{C}_0 \left(\|Hg\|_{L^2(\Omega)} + \|H\|_{L^\infty(\Omega)} C_\Omega \frac{L_1 C_\Omega + L_2}{c_0} \|g\|_{L^2(\Omega)} \right)$$

holds with

$$\tilde{C}_0 := \left(\frac{\beta + \|H\|_{L^\infty(\Omega)}(L_1 C_\Omega + L_2)}{c_0 \cdot \alpha} \right).$$

Proof. Due to (3.4), we know that there exist $\tilde{u}_{H,h}^{\mathrm{ms}} \in V_{H,h}^{\mathrm{ms}}$ and $\tilde{u}_h^{\mathrm{f}} \in V_h^{\mathrm{f}}$, such that

$$u_h = \tilde{u}_{H,h}^{\mathrm{ms}} + \tilde{u}_h^{\mathrm{f}}.$$

We use the Galerkin orthogonality obtained from the equations (3.1) and (3.7) to conclude for all $v \in V_{H,h}^{\mathrm{ms}}$,

$$\langle A\nabla(u_h - u_{H,h}^{\mathrm{ms}}), \nabla v\rangle + \langle F(u_h, \nabla u_h), v\rangle - \langle F\left(u_{H,h}^{\mathrm{ms}}, \nabla u_{H,h}^{\mathrm{ms}}\right), v\rangle = 0. \tag{4.1}$$

In particular $v = u_{H,h}^{\mathrm{ms}} - \tilde{u}_{H,h}^{\mathrm{ms}} \in V_{H,h}^{\mathrm{ms}}$ is an admissible test function in (4.1). Together with $\mathfrak{I}_H(\tilde{u}_h^{\mathrm{f}}) = 0$, this yields

$$\begin{aligned}
c_0|u_h - u_{H,h}^{\mathrm{ms}}|_{H^1(\Omega)}^2 &\overset{(2.2)}{\leq} \langle A\nabla(u_h - u_{H,h}^{\mathrm{ms}}), \nabla(u_h - u_{H,h}^{\mathrm{ms}})\rangle \\
&\quad + \langle F(u_h, \nabla u_h) - F(u_{H,h}^{\mathrm{ms}}, \nabla u_{H,h}^{\mathrm{ms}}), u_h - u_{H,h}^{\mathrm{ms}}\rangle \\
&\overset{(4.1)}{=} \langle A\nabla(u_h - u_{H,h}^{\mathrm{ms}}), \nabla(u_h - \tilde{u}_{H,h}^{\mathrm{ms}})\rangle \\
&\quad + \langle F(u_h, \nabla u_h) - F(u_{H,h}^{\mathrm{ms}}, \nabla u_{H,h}^{\mathrm{ms}}), u_h - \tilde{u}_{H,h}^{\mathrm{ms}}\rangle \\
&= \langle A\nabla(u_h - u_{H,h}^{\mathrm{ms}}), \nabla\tilde{u}_h^{\mathrm{f}}\rangle + \left\langle F(u_h, \nabla u_h) - F\left(u_{H,h}^{\mathrm{ms}}, \nabla u_h\right), \tilde{u}_h^{\mathrm{f}} - \mathfrak{I}_H(\tilde{u}_h^{\mathrm{f}})\right\rangle \\
&\quad + \left\langle F\left(u_{H,h}^{\mathrm{ms}}, \nabla u_h\right) - F\left(u_{H,h}^{\mathrm{ms}}, \nabla u_{H,h}^{\mathrm{ms}}\right), \tilde{u}_h^{\mathrm{f}} - \mathfrak{I}_H(\tilde{u}_h^{\mathrm{f}})\right\rangle \\
&\lesssim \beta|u_h - u_{H,h}^{\mathrm{ms}}|_{H^1(\Omega)}|\tilde{u}_h^{\mathrm{f}}|_{H^1(\Omega)} \\
&\quad + \|H\|_{L^\infty(\Omega)}(L_1\|u_h - u_{H,h}^{\mathrm{ms}}\|_{L^2(\Omega)} + L_2|u_h - u_{H,h}^{\mathrm{ms}}|_{H^1(\Omega)})|\tilde{u}_h^{\mathrm{f}}|_{H^1(\Omega)} \\
&\lesssim (\beta + \|H\|_{L^\infty(\Omega)}(L_1 C_\Omega + L_2)) \cdot |u_h - u_{H,h}^{\mathrm{ms}}|_{H^1(\Omega)} \cdot |\tilde{u}_h^{\mathrm{f}}|_{H^1(\Omega)}.
\end{aligned}$$

With $\langle A\nabla \tilde{u}_{H,h}^{\mathrm{ms}}, \nabla \tilde{u}_h^{\mathrm{f}}\rangle = 0$ and with $\mathfrak{I}_H(v_{\mathrm{f}}) = 0$ for all $v_{\mathrm{f}} \in V^{\mathrm{f}}$ we get

$$\begin{aligned}
\alpha|\tilde{u}_h^{\mathrm{f}}|^2_{H^1(\Omega)} &\leq \langle A\nabla \tilde{u}_h^{\mathrm{f}}, \nabla \tilde{u}_h^{\mathrm{f}}\rangle \\
&= \left\langle A\nabla u_h, \nabla \tilde{u}_h^{\mathrm{f}}\right\rangle = \left\langle g, \tilde{u}_h^{\mathrm{f}}\right\rangle - \left\langle F(u_h, \nabla u_h), \tilde{u}_h^{\mathrm{f}}\right\rangle \\
&= \left\langle g, \tilde{u}_h^{\mathrm{f}} - \mathfrak{I}_H(\tilde{u}_h^{\mathrm{f}})\right\rangle - \left\langle F(u_h, \nabla u_h), \tilde{u}_h^{\mathrm{f}} - \mathfrak{I}_H\left(\tilde{u}_h^{\mathrm{f}}\right)\right\rangle \\
&\overset{(2.3)}{\lesssim} \left(\|Hg\|_{L^2(\Omega)} + \|H\|_{L^\infty(\Omega)} C_\Omega \frac{L_1 C_\Omega + L_2}{c_0} \|g\|_{L^2(\Omega)}\right) \cdot |\tilde{u}_h^{\mathrm{f}}|_{H^1(\Omega)}.
\end{aligned}$$

The theorem follows by combing the results. □

The subsequent lemma is a consequence of the previous one.

Lemma 4.2. *Let $u_h \in V_h$ denote the fine scale approximation obtained from equation (3.1) and let $u_{H,h}^{\mathrm{ms},k} \in V_{H,h}^{\mathrm{ms},k}$ denote the solution of problem (3.11) (fully discrete LOD approximation). If the assumptions* (A1)−(A7) *hold true we obtain the estimate*

$$|u_h - u_{H,h}^{\mathrm{ms},k}|_{H^1(\Omega)} \lesssim \tilde{C}_2 \|g\|_{L^2(\Omega)} \|H\|_{L^\infty(\Omega)} + \tilde{C}_3 \min_{v_{H,h}^{\mathrm{ms},k} \in V_{H,h}^{\mathrm{ms},k}} \left\| A^{\frac{1}{2}} \nabla \left(u_{H,h}^{\mathrm{ms}} - v_{H,h}^{\mathrm{ms},k}\right)\right\|_{L^2(\Omega)},$$

where

$$\begin{aligned}
\tilde{C}_1 &:= (\beta + (L_1 C_\Omega + L_2) C_\Omega) \cdot \left(\frac{\beta + \|H\|_{L^\infty(\Omega)}(L_1 C_\Omega + L_2)}{c_0^2 \cdot \alpha}\right), \\
\tilde{C}_2 &:= \tilde{C}_1 + \tilde{C}_1 \cdot C_\Omega \frac{L_1 C_\Omega + L_2}{c_0}, \\
\tilde{C}_3 &:= \frac{\beta^{\frac{1}{2}} + \alpha^{-\frac{1}{2}}(L_1 C_\Omega + L_2) C_\Omega}{c_0}.
\end{aligned}$$

Proof. Let $v_{H,h}^{\mathrm{ms},k} \in V_{H,h}^{\mathrm{ms},k}$ denote an arbitrary element. Using the Galerkin orthogonality obtained from (3.1) and (3.11), we start in the same way as in the proof of Lemma 4.1 to get

$$\begin{aligned}
c_0|u_h - u_{H,h}^{\mathrm{ms},k}|^2_{H^1(\Omega)} &\\
\overset{(2.2)}{\leq}\ & \langle A\nabla(u_h - u_{H,h}^{\mathrm{ms},k}), \nabla(u_h - u_{H,h}^{\mathrm{ms},k})\rangle \\
&+\langle F(u_h, \nabla u_h) - F(u_{H,h}^{\mathrm{ms},k}, \nabla u_{H,h}^{\mathrm{ms},k}), u_h - u_{H,h}^{\mathrm{ms},k}\rangle \\
\overset{(4.1)}{=}\ & \langle A\nabla(u_h - u_{H,h}^{\mathrm{ms},k}), \nabla(u_h - u_{H,h}^{\mathrm{ms}}) + \nabla(u_{H,h}^{\mathrm{ms}} - v_{H,h}^{\mathrm{ms},k})\rangle \\
&+\langle F(u_h, \nabla u_h) - F(u_{H,h}^{\mathrm{ms},k}, \nabla u_{H,h}^{\mathrm{ms},k}), (u_h - u_{H,h}^{\mathrm{ms}}) + (u_{H,h}^{\mathrm{ms}} - v_{H,h}^{\mathrm{ms},k})\rangle \\
\leq\ & (\beta + (L_1 C_\Omega + L_2) C_\Omega)|u_h - u_{H,h}^{\mathrm{ms},k}|_{H^1(\Omega)} |u_h - u_{H,h}^{\mathrm{ms}}|_{H^1(\Omega)} \\
&+(\beta^{\frac{1}{2}} + \alpha^{-\frac{1}{2}}(L_1 C_\Omega + L_2) C_\Omega)|u_h - u_{H,h}^{\mathrm{ms},k}|_{H^1(\Omega)} \|A^{\frac{1}{2}} \nabla(u_{H,h}^{\mathrm{ms}} - v_{H,h}^{\mathrm{ms},k})\|_{L^2(\Omega)}.
\end{aligned}$$

Dividing by $|u_h - u_{H,h}^{\mathrm{ms},k}|_{H^1(\Omega)}$ and estimating $|u_h - u_{H,h}^{\mathrm{ms}}|_{H^1(\Omega)}$ with Lemma 4.1 yields the result. □

The combination of Lemmas 3.2 and 4.2 yields the main result of this paper.

Theorem 4.3. *Let $u_h \in V_h$ be solution of (3.1) and let $u_{H,h}^{\mathrm{ms},k} \in V_{H,h}^{\mathrm{ms},k}$ be the solution of (3.11). If the assumptions* (A1)−(A7) *hold true and if the number of layers k fulfills $k \gtrsim |\log(\|H\|_{L^\infty(\Omega)})|$, then it holds*

$$\left|u_h - u_{H,h}^{\mathrm{ms},k}\right|_{H^1(\Omega)} \lesssim \tilde{C}\|H\|_{L^\infty(\Omega)}\|g\|_{L^2(\Omega)},$$

where

$$\tilde{C} := \tilde{C}_2 + C_\Omega \frac{\beta}{c_0} \tilde{C}_3$$

and with $\tilde{C}_2$ *and* $\tilde{C}_3$ *as in Lemma 4.2.*

Proof. We define $w_{H,h}^{\mathrm{ms},k} \in V_{H,k}^{\mathrm{ms}}$ by

$$w_{H,h}^{\mathrm{ms},k} := \sum_{j=1}^{J} u_{H,h}^{\mathrm{ms}}(z_j)\lambda_{j,k}^{\mathrm{ms}} = \sum_{j=1}^{J} u_{H,h}^{\mathrm{ms}}(z_j)\left(\lambda_j - \phi_{j,k}^h\right)$$

where $u_{H,h}^{\mathrm{ms}}(z_j)$, $j = 1, 2, \dots, J$, are the coefficients in the basis representation of $u_{H,h}^{\mathrm{ms}}$ from Definition 3.1. Hence,

$$\begin{aligned} &\min_{v_{H,h}^{\mathrm{ms},k} \in V_{H,h}^{\mathrm{ms},k}} \left\| A^{\frac{1}{2}} \nabla \left(u_{H,h}^{\mathrm{ms}} - v_{H,h}^{\mathrm{ms},k} \right) \right\|_{L^2(\Omega)}^2 \\ &\leq \left\| A^{\frac{1}{2}} \nabla \left(u_{H,h}^{\mathrm{ms}} - w_{H,h}^{\mathrm{ms},k} \right) \right\|_{L^2(\Omega)}^2 \\ &\lesssim \sum_{j=1}^{J} k^d u_{H,h}^{\mathrm{ms}}(z_j)^2 \| A^{1/2} \nabla \left(\phi_j^h - \phi_{j,k}^h \right) \|_{L^2(\Omega)}^2. \end{aligned} \tag{4.2}$$

For details on the last step, we refer to Lemma 4.9 in [31]. Due to the Galerkin orthogonality for the corrector problems it is possible to show

$$\| A^{1/2} \nabla (\phi_j^h - \phi_{j,k}^h) \|_{L^2(\Omega)}^2 \lesssim \| A^{1/2} \nabla \phi_j^h \|_{L^2(\Omega \setminus \omega_{j,k-1})}^2, \tag{4.3}$$

where the idea behind the proof of (4.3) is to use the best approximation property of $\phi_{j,k}^h$ in $V_h^{\mathrm{f}}(\omega_{j,k})$ to replace it by an arbitrary other function from $V_h^{\mathrm{f}}(\omega_{j,k})$. The best choice would be $\mathbb{1}_{\omega_{j,k}} \phi_j^h$, where $\mathbb{1}_{\omega_{j,k}}$ is the indicator function of $\omega_{j,k}$ (this choice would directly give the result). However, $\mathbb{1}_{\omega_{j,k}} \phi_j^h$ is not in $V_h^{\mathrm{f}}(\omega_{j,k})$, which is why additional interpolation and projection operators are required. The rather technical details for the proof of (4.3) are therefore given in the first part of the proof of Lemma 4.8 in [31].

The application of Lemma 3.2, (3.5), (4.3) and some inverse inequality yield

$$\begin{aligned} \| A^{1/2} \nabla (\phi_j^h - \phi_{j,k}^h) \|_{L^2(\Omega)}^2 &\lesssim \mathrm{e}^{-2rk} \| A^{1/2} \nabla \phi_j^h \|_{L^2(\Omega)}^2 \\ &\leq \mathrm{e}^{-2rk} \| A^{1/2} \nabla \lambda_j \|_{L^2(\Omega)}^2 \\ &\leq \beta \mathrm{e}^{-2rk} \| H \|_\infty^{-2} \| \lambda_j \|_{L^2(\Omega)}^2, \end{aligned}$$

with a generic rate r that is proportional to $(\beta/\alpha)^{1/2}$. By choosing $k = m \cdot |\log(\|H\|_{L^\infty(\Omega)})|$ with $m \in \mathbb{N}$, we can achieve an arbitrary fast polynomial convergence of this term in H (this will also cancel the k^d term). However, we bound this by a linear convergence since this is fastest rate that we can obtain for the whole error. Finally, the combination of this estimate and (4.2) plus

$$\begin{aligned} \sum_{j=1}^{J} u_{H,h}^{\mathrm{ms}}(z_j)^2 \| \lambda_j \|_{L^2(\Omega)}^2 &\lesssim \left\| \sum_{j=1}^{J} u_{H,h}^{\mathrm{ms}}(z_j) \lambda_j \right\|_{L^2(\Omega)}^2 \\ &= \left\| \sum_{j=1}^{J} u_{H,h}^{\mathrm{ms}}(z_j) ((\mathfrak{I}_H|_{V_H})^{-1} \circ \mathfrak{I}_H)(\lambda_j - \phi_j^h) \right\|_{L^2(\Omega)}^2 \\ &= \| (\mathfrak{I}_H|_{V_H})^{-1} \circ \mathfrak{I}_H) u_{H,h}^{\mathrm{ms}} \|_{L^2(\Omega)}^2 \overset{(\mathrm{A5})+(\mathrm{A6})}{\lesssim} \| \nabla u_{H,h}^{\mathrm{ms}} \|_{L^2(\Omega)}^2 \leq C_\Omega^2 c_0^{-2} \| g \|_{L^2(\Omega)}^2 \end{aligned}$$

yields the assertion. □

5. The multiscale newton scheme

In this section we discuss a solution algorithm for handling the nonlinear multiscale problem (3.11). For this purpose, we consider a damped Newton's method in the multiscale space $V_{H,h}^{\mathrm{ms},k}$. Recall that we are looking for $u \in H_0^1(\Omega)$ with

$$\langle B(u), v\rangle_{H^{-1},H_0^1} = \langle g, v\rangle \quad \text{for all } v \in H_0^1(\Omega),$$

where we introduced the notation

$$\langle B(v), w\rangle_{H^{-1},H_0^1} := \langle A\nabla v, \nabla w\rangle + \langle F(\cdot, v, \nabla v), w\rangle.$$

Here, $B : H_0^1(\Omega) \to H^{-1}(\Omega)$ is a hemicontinuous and strongly monotone operator due to assumption (A3). As already mentioned, under these assumptions, the Browder−Minty theorem yields a unique solution of the above problem. However, we will need an additional assumption on F to guarantee that the Newton scheme converges.

Assumption 3. Let $DF(x,\cdot,\cdot)$ denote the Jacobian matrix of $F(x,\cdot,\cdot)$.

(A8) We assume that there exists some constant $L_D \geq 0$ so that for almost every x in Ω and for all $(\xi_1,\zeta_1) \in \mathbb{R}\times\mathbb{R}^d$ and $(\xi_2,\zeta_2) \in \mathbb{R}\times\mathbb{R}^d$

$$|DF(x,\xi_1,\zeta_1) - DF(x,\xi_2,\zeta_2)| \leq L_D|(\xi_1,\zeta_1) - (\xi_2,\zeta_2)|,$$

i.e. $F(x,\cdot,\cdot) \in W^{2,\infty}(\mathbb{R}\times\mathbb{R}^d)$.

For clarity of the presentation we will leave out several indices within this section. In particular, we make use of the following notation.

Definition 5.1. For simplicity, we define

$$V^{\mathrm{ms}} := V_{H,h}^{\mathrm{ms},k} \quad \text{with basis } \lambda_j^{\mathrm{ms}} := \lambda_{j,k}^{\mathrm{ms}} = \lambda_j - \phi_{j,k}^h \text{ for } 1 \leq j \leq J.$$

Furthermore, we denote $u^{\mathrm{ms}} := u_{H,h}^{\mathrm{ms},k}$. Additionally, let

$$\partial_1 F(x,\xi,\zeta) := \partial_\xi F(x,\xi,\zeta) \quad \text{and} \quad \partial_2 F(x,\xi,\zeta) := \partial_\zeta F(x,\xi,\zeta).$$

We now describe the Newton strategy in detail. The fully discrete multiscale problem is to

$$\text{find } u^{\mathrm{ms}} \in V^{\mathrm{ms}} : \quad \langle A\nabla u^{\mathrm{ms}}, \nabla\lambda_j^{\mathrm{ms}}\rangle + \langle F(\cdot, u^{\mathrm{ms}}, \nabla u^{\mathrm{ms}}), \lambda_j^{\mathrm{ms}}\rangle - \langle g, \lambda_j^{\mathrm{ms}}\rangle = 0$$

for all $1 \leq j \leq J$. Again, using Browder−Minty, u^{ms} exists and is unique. Accordingly, we get the following well posed algebraic version of the problem:

$$\text{find } \bar{\alpha} \in \mathbb{R}^J : \quad G(\bar{\alpha}) = 0$$

and where $G : \mathbb{R}^J \to \mathbb{R}^J$ is given by

$$(G(\alpha))_l := \sum_{j=1}^J \alpha_j \langle A\nabla\lambda_j^{\mathrm{ms}}, \nabla\lambda_l^{\mathrm{ms}}\rangle + \langle F\left(\cdot, \sum_{j=1}^J \alpha_j\lambda_j^{\mathrm{ms}}, \sum_{j=1}^J \alpha_j\nabla\lambda_j^{\mathrm{ms}}\right), \lambda_l^{\mathrm{ms}}\rangle - \langle g, \lambda_l^{\mathrm{ms}}\rangle. \tag{5.1}$$

We have the relation $u^{\mathrm{ms}} = \sum_{j=1}^J \bar{\alpha}_j\lambda_j^{\mathrm{ms}}$. Before we can apply the Newton method to (5.1), we need to ensure that the iterations of the scheme are well defined.

Lemma 5.2. *Let $(X, \|\cdot\|_X)$ denote a Hilbert space with dual space X'. Let furthermore $B : X \to X'$ be a hemicontinuous, Fréchet differentiable and strongly monotone operator on X,* i.e. *there exists $c_0 > 0$ so that*

$$\begin{aligned}&\langle B(v) - B(w), v - w\rangle_X \geq c_0\|v - w\|_X^2 \quad \text{for all } v, w \in X \text{ and}\\ &\quad s \mapsto \langle B(u + sv), w\rangle_X\end{aligned}$$

is a continuous function on $[0, 1]$ for all $u, v, w \in X$. Let X_N denote a finite dimensional subspace with basis $\{\psi_1, \dots, \psi_N\}$ and let $b : \mathbb{R}^N \to V_N$ define the linear bijection with $b(\alpha) := \sum_{i=1}^N \alpha_i\psi_i$. If $G(\alpha) := b^{-1}(B(b(\alpha)))$, then the Jacobi matrix $DG(\alpha) \in \mathbb{R}^{n\times n}$ has only positive eigenvalues.

Proof. Let B' denote the Fréchet derivative of B, given by

$$B'(u)(v) = \lim_{s\to 0} \frac{B(u + sv) - B(u)}{s} \quad \text{for } u, v \in X.$$

This and the strong monotonicity yield

$$\begin{aligned}\langle B'(u)(v), v\rangle_{H^{-1},H_0^1} &= \lim_{s\to 0} \frac{(B(u + sv) - B(u))(v)}{s}\\ &= \lim_{s\to 0} \frac{1}{s^2}(B(u + sv) - B(u))(u + sv - u)\\ &\geq \lim_{s\to 0} \frac{1}{s^2} c_0\|sv\|^2 = c_0\|v\|^2.\end{aligned} \tag{5.2}$$

Next, observe that b induces an inner product on $\mathbb{R}^N$ by $(\alpha_1, \alpha_2)_b := \langle b(\alpha_1), b(\alpha_2)\rangle_X$. Let $\alpha := b^{-1}(u)$ then we get

$$\begin{aligned}B'(u)(\psi_i) &= \lim_{s\to 0} \frac{B(u + s\psi_i) - B(u)}{s}\\ &= \lim_{s\to 0} \frac{(b \circ b^{-1})\left(B\left(\sum_{j=1}^N (\alpha_j + s\delta_{ij})\,\psi_j\right) - (b \circ b^{-1})\left(B\left(\sum_{j=1}^N \alpha_j\psi_j\right)\right)\right.}{s}\\ &= b\left(\lim_{s\to 0} \frac{G(\alpha + se_i) - G(\alpha)}{s}\right)\\ &= b(D_\alpha G(\alpha)e_i).\end{aligned}$$

Using this, we get for arbitrary $\xi \in \mathbb{R}^N$ and $v_\xi := b(\xi)$,

$$\begin{aligned}(D_\alpha G(\alpha)\xi, \xi)_b &= \sum_{i,j}^N \xi_i\xi_j(D_\alpha G(\alpha)e_i, e_j)_b\\ &= \sum_{i,j}^N \xi_i\xi_j(b(D_\alpha G(\alpha)e_i), b(e_j))_X\\ &= \sum_{i,j}^N \xi_i\xi_j(B'(u)(\psi_i), \psi_j)_X\\ &= (B'(u)(v_\xi), v_\xi)_X \overset{(5.2)}{\geq} c_0\|v_\xi\|_X^2 = c_0\|\xi\|_b^2.\end{aligned}$$

Since all norms in $\mathbb{R}^N$ are equivalent we have the desired result. □

Now, we can apply the Newton method for solving the nonlinear algebraic equation $G(\bar{\alpha}) = 0$. If $D_\alpha G$ denotes the Jacobian matrix of G, we get the following iteration scheme:

$$\alpha^{(n+1)} := \alpha^{(n)} + \triangle\alpha^{(n)},$$

where $\triangle\alpha^{(n)}$ solves

$$D_\alpha G\left(\alpha^{(n)}\right) \triangle\alpha^{(n)} = -G\left(\alpha^{(n)}\right). \tag{5.3}$$

Here, $D_\alpha G$ is given by

$$D_{\alpha_i}\left(G(\alpha)\right)_l := \langle A\nabla\lambda_i^{\mathrm{ms}}, \nabla\lambda_l^{\mathrm{ms}}\rangle + \left\langle \partial_1 F\left(\cdot, \sum_{j=1}^{J}\alpha_j\lambda_j^{\mathrm{ms}}, \sum_{j=1}^{J}\alpha_j\nabla\lambda_j^{\mathrm{ms}}\right)\lambda_i^{\mathrm{ms}}, \lambda_l^{\mathrm{ms}}\right\rangle$$
$$+ \left\langle \partial_2 F\left(\cdot, \sum_{j=1}^{J}\alpha_j\lambda_j^{\mathrm{ms}}, \sum_{j=1}^{J}\alpha_j\nabla\lambda_j^{\mathrm{ms}}\right)\cdot\nabla\lambda_i^{\mathrm{ms}}, \lambda_l^{\mathrm{ms}}\right\rangle.$$

Lemma 5.2 ensures that equation (5.3) has a unique solution $\triangle\alpha^{(n)}$, *i.e.* that the Newton iteration is well posed. Since $G \in C^1(\mathbb{R}^N)$ has a nonsingular Jacobian matrix $D_\alpha G$ (due to Lem. 5.2) and since we have Lipschitz-continuity of $D_\alpha G$ (due to Assumption 3), we have that the Newton scheme converges quadratically as long as the starting value is close enough to the exact solution (*cf.* [12]). However, this means that we can only guarantee local convergence of the method. In order to ensure global convergence, we can use a simple damping strategy due to Armijo [2]. Here we are looking for a damping parameter $\zeta \in (0, 1]$ so that $\alpha^{(n+1)} := \alpha^{(n)} + \zeta\triangle\alpha^{(n)}$ with the property $|G(\alpha^{(n+1)})| < (1 - \frac{\zeta}{2})|G(\alpha^{(n)})|$. In our case, the convergence of the damped Newton scheme can be guaranteed by the following lemma which is based on the results by Kelley [26].

Lemma 5.3. *Let assumptions* (A1)−(A3) *and* (A8) *be fulfilled, then the damped Newton scheme converges,* i.e. *there exists a nonempty (damping) interval* $[\zeta_0, \zeta_1] \subset (0, 1)$*, so that*

$$\left|G\left(\alpha^{(n+1)}\right)\right| < \left(1 - \frac{\zeta}{2}\right)\left|G\left(\alpha^{(n)}\right)\right| \qquad \text{for all } \zeta \in [\zeta_0, \zeta_1].$$

Here, $\zeta_0 > 0$ *is independent of* $\alpha^{(n)}$ *and* $\triangle\alpha^{(n)}$*, which prevents* $\zeta_1 \to 0$*.*

Proof. The existence of a damping parameter so that $|G(\alpha^{(n+1)})| < |G(\alpha^{(n)})|$ is an easy observation if we look at the function $h(\zeta) := |G(\alpha^{(n)} + \zeta\triangle\alpha^{(n)})|^2$ which fulfills $h(0) > 0$ and $h'(0) = -2G(\alpha^{(n)}) \cdot G(\alpha^{(n)}) < 0$. The existence of a uniform lower bound $\zeta_0 > 0$ was proved by Kelley ([26], Lem. 8.2.1 and Thm. 8.2.1 therewithin). The results by Kelley require Lipschitz continuity of $D_\alpha G$ (guaranteed by Assump. (A8)) and uniform boundedness of $|(D_\alpha G(\alpha))^{-1}|$. The latter one is fulfilled since the proof of Lemma 5.2 shows that the smallest eigenvalue of $(D_\alpha G(\alpha))$ is equal or larger than c_0. This implies that the largest eigenvalue of $(D_\alpha G(\alpha))^{-1}$ is bounded by c_0^{-1}, hence $|(D_\alpha G(\alpha))^{-1}|$ is uniformly bounded. □

In summary, Lemma 5.3 guarantees globally linear convergence of the method (using damping) and locally (*i.e.* in an environment of the solution) even quadratic convergence using the classical Newton scheme without damping. With these considerations, we can state the full algorithm below. Recall that $\mathcal{N}_H$ denotes the set of interior vertices of $\mathcal{T}_H$ and for $z_j \in \mathcal{N}_H$, $\lambda_j \in V_H$ denotes the corresponding nodal basis function.

Note that in the presented algorithm, each iteration starts with the damping parameter $\zeta_n = 1$ and we do not use damping parameters from previous iterations. The advantage is that we automatically get quadratic convergence of the Newton scheme as soon as we leave the region where damping is required. Therefore, damping is only used when really necessary.

Algorithm: dampedNewtonLOD(*abstol*, *reltol*, $\alpha^{(0)}$, k)

In parallel **foreach** $z_j \in \mathcal{N}_H$ **do**
compute $\phi_{j,k}^h \in V_h^{\mathrm{f}}(\omega_{j,k})$ with

$$\langle A\nabla\phi_{j,k}^h, \nabla w\rangle = \langle A\nabla\lambda_j, \nabla w\rangle \quad \text{for all } w \in V_h^{\mathrm{f}}(\omega_{j,k}).$$

end
Set $V_{H,h}^{\mathrm{ms},k} := \operatorname{span}\{\lambda_j - \phi_{j,k}^h \mid 1 \le j \le J\}$. Set $\lambda_{j,k}^{\mathrm{ms}} = \lambda_j - \phi_{j,k}^h$.

Set $\alpha^{(n)} := \alpha^{(0)}$. Set $u_{H,h}^{\mathrm{ms},k,(n)} := \sum_{j=1}^J \alpha_j^{(n)} \lambda_{j,k}^{\mathrm{ms}}$. Set

$$(G(\alpha))_i := \sum_{j=1}^J \alpha_j \langle A\nabla\lambda_{j,k}^{\mathrm{ms}}, \nabla\lambda_{i,k}^{\mathrm{ms}}\rangle + \langle F(\cdot, \sum_{j=1}^J \alpha_j \lambda_{j,k}^{\mathrm{ms}}, \sum_{j=1}^J \alpha_j \nabla\lambda_{j,k}^{\mathrm{ms}}) - g, \lambda_{i,k}^{\mathrm{ms}}\rangle.$$

Set $tol := |G(\alpha^{(0)})|_2 \cdot reltol + abstol$.

while $|G(\alpha^{(n)})|_2 > tol$ **do**
Set $u_{H,h}^{\mathrm{ms},k,(n)} := \sum_{j=1}^J \alpha_j^{(n)} \lambda_{j,k}^{\mathrm{ms}}$.
Define the entries of the stiffness matrix $M^{(n)}$ by

$$M_{il}^{(n)} := \langle A\nabla\lambda_{l,k}^{\mathrm{ms}}, \nabla\lambda_{i,k}^{\mathrm{ms}}\rangle + \langle \partial_1 F(\cdot, u_{H,h}^{\mathrm{ms},k,(n)}, \nabla u_{H,h}^{\mathrm{ms},k,(n)})\lambda_{l,k}^{\mathrm{ms}}, \lambda_{i,k}^{\mathrm{ms}}\rangle + \langle \partial_2 F(\cdot, u_{H,h}^{\mathrm{ms},k,(n)}, \nabla u_{H,h}^{\mathrm{ms},k,(n)}) \cdot \nabla\lambda_{l,k}^{\mathrm{ms}}, \lambda_{i,k}^{\mathrm{ms}}\rangle.$$

Define the entries of the right hand side by

$$F_i^{(n)} := \langle g, \lambda_{i,k}^{\mathrm{ms}}\rangle - \langle A\nabla u_{H,h}^{\mathrm{ms},k,(n)}, \nabla\lambda_{i,k}^{\mathrm{ms}}\rangle - \langle F(\cdot, u_{H,h}^{\mathrm{ms},k,(n)}, \nabla u_{H,h}^{\mathrm{ms},k,(n)}), \lambda_{i,k}^{\mathrm{ms}}\rangle.$$

Find $(\triangle\alpha)^{(n+1)} \in \mathbb{R}^J$, with

$$M^{(n)}(\triangle\alpha)^{(n+1)} = F^{(n)}.$$

Set $\zeta_n := 1$. Set $\alpha^{(n+1)} := \alpha^{(n)} + \zeta_n \triangle\alpha^{(n)}$.
while $|G(\alpha^{(n+1)})| \ge (1 - \frac{\zeta_n}{2})|G(\alpha^{(n)})|$ **do**
Set $\zeta_n := \frac{1}{2}\zeta_n$. Set $\alpha^{(n+1)} := \alpha^{(n)} + \zeta_n \triangle\alpha^{(n)}$.
end
Set $\alpha^{(n)} := \alpha^{(n+1)}$. Set $tol := |G(\alpha^{(n)})|_2 \cdot reltol + abstol$.
end

Set $u_{H,h}^{\mathrm{ms},k,(n)} := \sum_{j=1}^J \alpha_j^{(n)} \lambda_{j,k}^{\mathrm{ms}}$.

Proposition 5.4. *We use the notation stated in Definition 5.1. Let $u \in H_0^1(\Omega)$ denote the solution of (2.1), let $u_h \in V_h$ denote the solution of (3.1) and let $u^{\mathrm{ms}} \in V^{\mathrm{ms}}$ denote the solution of (3.11). Furthermore, we let $u^{\mathrm{ms},(n)} := u_{H,h}^{\mathrm{ms},k,(n)}$ define the n'th iterate from the damped Newton LOD Method stated in the algorithm. Under assumptions* (A1)–(A8)*, the Newton step (5.3) is well posed, yields an unique solution and $u^{\mathrm{ms},(n)}$ converges at least linearly to u^{ms}. If furthermore $k \gtrsim |\log(\|H\|_{L^\infty(\Omega)})|$, the* a priori *error estimate*

$$\|u - u^{\mathrm{ms}}\|_{H^1(\Omega)} \le C\left(\|H\|_{L^\infty(\Omega)} + \|u - u_h\|_{H^1(\Omega)}\right)$$

holds with a generic constant $C = O(1)$ (see Thms. 3.5 and 4.3 for details) and

$$\left\|u^{\mathrm{ms}} - u^{\mathrm{ms},(n)}\right\|_{H^1(\Omega)} \le L_n(H)\left\|u^{\mathrm{ms}} - u^{\mathrm{ms},(n-1)}\right\|_{H^1(\Omega)}.$$

Here, we have $L_n(H) < 1$.

If $u^{\mathrm{ms},(n-1)}$ is sufficiently close to u^{ms}, we even get quadratic convergence of the Newton scheme, i.e.,

$$\left\|u^{\mathrm{ms}} - u^{\mathrm{ms},(n)}\right\|_{H^1(\Omega)} \leq L_n(H) \left\|u^{\mathrm{ms}} - u^{\mathrm{ms},(n-1)}\right\|^2_{H^1(\Omega)}.$$

with

$$L_n(H) \leq \frac{\|(D_\alpha G)^{-1}\|_{L^\infty(\mathbb{R}^N)}}{L},$$

where L denotes the Lipschitz-constant of $D_\alpha G$. As indicated, $L_n(H)$ typically depends on the mesh size. However, in some cases of semi-linear problems, it is possible to bound $L_n(H)$ independent of the triangulation (cf. [25]*). In particular, if $F(x,u,\nabla u) = F(x,u)$ (*i.e. *no dependency on ∇u) we get that $L_n(H) = L_n$ independent of the underlying mesh. The proof can be obtained analogously to the proof of Proposition 4.1 in* [25]*. The proof fails for general $F(x,u,\nabla u)$.*

Remark 5.5. Note that the proposed method only requires the computation of the multiscale basis $\{\lambda_j^{\mathrm{ms}} | \, 1 \leq j \leq J\}$ once at the beginning. For each iteration step of the damped Newton scheme, (5.3) is a low dimensional linear problem that can reuse the initially computed multiscale basis. If the multiscale basis was computed using the nonlinear term F, local corrector problems would have to be solved for each Newton step newly, making the whole procedure significantly more expensive. We also note that assemblation of the tangent matrix $M^{(n)}$ and the residual $F^{(n)}$ still requires a quadrature rule that captures the fine scale features. Depending on the type of the nonlinearity this might have to be done newly for each iteration step, making the quadrature rule a significant part of each Newton step.

6. Numerical experiment

As mentioned in the introduction, Richards-type equations can be an application of our LOD-Newton framework. In general, the stationary Richards equation cannot necessarily be described by a monotone operator, however depending on the chosen model and the considered hydrological effects (including hysteresis, root uptake, friction, reaction fronts, *etc.*) monotone operators can arise in certain applications. One explicit example is the (regularized) time-discretized Kirchhoff transformed Richards equation regarded in [6]. For the case that there is no Signorini boundary condition prescribed, the problem that has to be solved for each time step corresponds to a nonlinear elliptic monotone problem (on the full space) that also fulfills the required assumption of Lipschitz-continuity.

Let us now consider the stationary Kirchhoff-transformed Richards equation

$$\nabla \cdot (K \nabla u) - \nabla \cdot (K \, kr(M(u)) \vec{g}) = f, \tag{6.1}$$

where u denotes the generalized pressure, K the hydraulic conductivity and kr the relative permeability depending on the saturation. kr is a monotone increasing function with values between 0 and 1 (typically bounded away from 0 to avoid degeneracy). If we have already full saturation, water cannot be conducted anymore, if the soil is completely dry (saturation is zero), water can be perfectly conducted. Formulas for kr were *e.g.* provided by Burdine [9] and Mualem [32]. In applications the variations of the hydraulic conductivity K are assumed to be constant (or at least slow) in gravity direction $\vec{g} = (0,0,\vec{g}_z)$, where $\vec{g}_z$ denotes the gravity factor of $9.81 m/s^2$. Soil probes are often only taken once in vertical direction, but a lot of samples are required to describe the variations of conductivity in horizontal direction. As a reduction of complexity one can often assume that $\nabla \cdot (K\vec{g}) = \partial_z(K_{zz}\vec{g}_z) = 0$ to consider the reduced equation

$$\nabla \cdot (K \nabla u) - \, (kr \circ M)'(u) \, (K\vec{g}) \cdot \nabla u = f. \tag{6.2}$$

Here we have $M(u) := \theta \circ \kappa^{-1}$, where θ denotes the saturation (depending on the pressure) and κ^{-1} the inverse of the Kirchhoff transformation $\kappa(p) := \int_0^p kr(\theta(q))\, dq$. The saturation θ can be obtained by the capillary pressure

TABLE 1. Results for fine grid with $\epsilon > h = 2^{-6} \approx 0.016 > \epsilon^{\frac{3}{2}}$ which resolves the oscillations of the linear term, but not the oscillations of the nonlinear term. The truncation parameter k determines the patch size by (3.8). We observe an average EOC of 2.37 for the L^2-error and an average EOC of 1.33 for the H^1-error.

H	k	$\left\|u_{H,h}^{\mathrm{ms},k} - u_h\right\|_{L^2(\Omega)}$	$\left\|u_{H,h}^{\mathrm{ms},k} - u_h\right\|_{H^1(\Omega)}$
2^{-2}	1	0.1455	1.6985
2^{-3}	2	0.0097	0.3737
2^{-4}	3	0.0023	0.1772
2^{-5}	3	0.0008	0.1067

relation (soil-water retention curves). Various explicit formulas for θ are available, see *e.g.* Van Genuchten [24], Brooks-Corey [8] or the Gardner model [23]. Depending on the chosen model $(kr \circ M)'$ might not be a Lipschitz continuous function, still regularization is possible. In the following numerical experiment, we consider a test problem that has the structure derived from a regularized Burdine–Brooks–Corey model. The corresponding explicit formulas for $(kr \circ M)$ are taken from [3]. Contrary to the model (6.2), we use a nonlinear advection term that is faster oscillating than the diffusion term. The reason is that we want to emphasize our claim, that the oscillations of the nonlinearity F do in fact not influence the convergence. Before stating the test problem related to (6.2), let us note that the method and the analytical results of this paper directly transfer to equations in divergence form like (6.1), *i.e.* the gradient in the weak formulation can be on the test function, as long as $F(x,u)$ does not dependent on the gradient ∇u.

We consider the following nonlinear advection-diffusion problem. Let $\Omega :=]0,1[^2$ and $\epsilon := 0.05$. Find u^ϵ with

$$\begin{aligned} -\nabla \cdot (A^\epsilon(x)\nabla u^\epsilon(x)) + \frac{1}{2}F^\epsilon(x,u^\epsilon)\partial_{x_2}u^\epsilon(x) &= -\frac{3}{10} \quad \text{in } \Omega \\ u^\epsilon(x) &= 0 \qquad \text{on } \partial\Omega, \end{aligned}$$

where A^ϵ is given by

$$A^\epsilon(x_1,x_2) := \frac{1}{8\pi^2}\begin{pmatrix} 2\left(2+\cos\left(2\pi\frac{x_1}{\epsilon}\right)\right)^{-1} & 0 \\ 0 & 1+\frac{1}{2}\cos\left(2\pi\frac{x_1}{\epsilon}\right) \end{pmatrix}$$

and

$$F^\epsilon(x,u) := \frac{1}{8\pi^2}\left(2+\cos\left(2\pi\frac{x_1}{\epsilon^{\frac{3}{2}}}\right)\right)\begin{cases} \sqrt{\frac{u}{2}+\frac{3}{2}} & \text{for } -3 \le u \le -\frac{5}{4} \\ p(u) & \text{for } -\frac{5}{4} \le u \le -1 \\ 0 & \text{for } u \ge -1 \end{cases},$$

where $p(u) = au^3 + bu^2 + cu + d$ is such that $F^\epsilon(x,\cdot) \in C^1(-3,\infty)$ for all $x \in \Omega$. The (unknown) exact solution of this problem takes values between 0 and -1.75.

The numerical experiments presented in this section were performed with a little different implementation of the localization strategy than the one described in Section 3.4. We used the localized basis functions proposed in [19], which have the completely same analytical properties than (3.9)–(3.10), with the only difference that they are computed with respect to unit vectors instead of gradients of basis functions in order to slightly stabilize the computations.

The tolerance *tol* in the Newton algorithm is set to 10^{-10}. We keep the resolution of the (uniformly refined) fine grid fixed with $h = 2^{-6} < \epsilon$. The computations were made for four different coarse grid resolutions $H = 2^{-2}, \ldots, 2^{-5}$.

For given H, we guess the truncation parameter k (according to (3.8)) by $|\log(H)|$. By log we mean the logarithm to the basis e. For $H = 2^{-l}$, $l = 2, \ldots, 5$ we obtain $\log(4) \approx 1.386$, $\log(8) \approx 2.08$, $\log(16) \approx 2.77$ and $\log(32) \approx 3.47$. Optimistically rounding we set the truncation parameter k to 1 for $H = 2^{-2}$, 2 for $H = 2^{-3}$, 3 for $H = 2^{-4}$ and 3 for $H = 2^{-5}$. The corresponding results are depicted in Table 1. We observe that the proportionality coefficient in the choice of the diameter of the patches $\mathrm{O}(\mathrm{diam}(\omega_{j,k})) \sim H|\log(\|H\|_{L^\infty(\Omega)})|$ can chosen to be on 1 without suffering from pre-asymptotic effects. In fact, we obtain an experimental order of convergence (EOC) of 2.37 for the L^2-error and an EOC of 1.33 for the H^1-error. The patches remain small and computational demand for solving the local problems remains very small. For further numerical studies of the method and the choice of patch sizes in the linear case, we refer to [31].

Acknowledgements. We would like to thank the anonymous reviewers for their valuable suggestions and their constructive criticism that helped us to improve the paper.

References

[1] H.W. Alt and S. Luckhaus, Quasilinear elliptic-parabolic differential equations. *Math. Z.* **183** (1983) 311–341.

[2] L. Armijo, Minimization of functions having Lipschitz continuous first partial derivatives. *Pacific J. Math.* **16** (1966) 1–3.

[3] H. Berninger, Domain Decomposition Methods for Elliptic Problems with Jumping Nonlinearities and Application to the Richards Equation. *Ph.D. thesis.* Freie Universität Berlin (2007).

[4] H. Berninger, Non-overlapping domain decomposition for the Richards equation *via* superposition operators. Vol. 70 of *Lect. Notes Comput. Sci. Eng.* Springer, Berlin (2009) 169–176.

[5] H. Berninger, R. Kornhuber and O. Sander, On nonlinear Dirichlet-Neumann algorithms for jumping nonlinearities. Domain decomposition methods in science and engineering XVI. Vol. 55 of *Lect. Notes Comput. Sci. Eng.* Springer, Berlin (2007) 489–496.

[6] H. Berninger, R. Kornhuber and O. Sander, Fast and robust numerical solution of the Richards equation in homogeneous soil. *SIAM J. Numer. Anal.* **49** (2011) 2576–2597.

[7] A. Bourlioux and A.J. Majda, An elementary model for the validation of flamelet approximations in non-premixed turbulent combustion. *Combust. Theory Model.* **4** (2000) 189–210.

[8] R.H. Brooks and A.T. Corey, Hydraulic properties of porous media. *Hydrol. Pap.* 4, Colo. State Univ., Fort Collins (1964).

[9] N.T. Burdine, Relative permeability calculations from pore-size distribution data. *Petr. Trans. Am. Inst. Mining Metall. Eng.* **198** (1953) 71–77.

[10] C. Carstensen, Quasi-interpolation and *a posteriori* error analysis in finite element methods. *ESAIM: M2AN* **33** (1999) 1187–1202.

[11] C. Carstensen and R. Verfürth, Edge residuals dominate *a posteriori* error estimates for low order finite element methods. *SIAM J. Numer. Anal.* **36** (1999) 1571–1587.

[12] J.E. Dennis Jr. and R.B. Schnabel, Numerical Methods for Unconstrained Optimization and Nonlinear Equations. *SIAM Classics Appl. Math.* (1996).

[13] W. E and B. Engquist, The heterogeneous multiscale methods. *Commun. Math. Sci.* **1** (2003) 87–132.

[14] A. Gloria, An analytical framework for the numerical homogenization of monotone elliptic operators and quasiconvex energies. *SIAM Multiscale Model. Simul.* **5** (2006) 996–1043.

[15] P. Henning, Convergence of MsFEM approximations for elliptic, non-periodic homogenization problems. *Netw. Heterog. Media* **7** (2012) 503–524.

[16] P. Henning and M. Ohlberger, The heterogeneous multiscale finite element method for advection-diffusion problems with rapidly oscillating coefficients and large expected drift. *Netw. Heterog. Media* **5** (2010) 711–744.

[17] P. Henning and M. Ohlberger, A Note on Homogenization of Advection-Diffusion Problems with Large Expected Drift. *Z. Anal. Anwend.* **30** (2011) 319–339.

[18] P. Henning and M. Ohlberger, Error control and adaptivity for heterogeneous multiscale approximations of nonlinear monotone problems. Preprint 01/11 – N, to appear in *DCDS-S, special issue on Numerical Methods based on Homogenization and Two-Scale Convergence* (2011).

[19] P. Henning and D. Peterseim, Oversampling for the Multiscale Finite Element Method. *SIAM Multiscale Model. Simul.* **12** (2013) 1149–1175.

[20] T. Hou and X.-H. Wu, A multiscale finite element method for elliptic problems in composite materials and porous media. *J. Comput. Phys.* **134** (1997) 169–189.

[21] 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) 3–24.

[22] T.J.R. Hughes and G. Sangalli, Variational multiscale analysis: the fine-scale Green?s function, projection, optimization, localization, and stabilized methods. *SIAM J. Numer. Anal.* **45** (2007) 539–557.

[23] W.R. Gardner, Some steady state solutions of unsaturated moisture ßow equations with application to evaporation from a water table. *Soil Sci.* **85** (1958) 228–232.

[24] M.T. van Genuchten, A closedform equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Sci. Soc. Am. J.* **44** (1980) 892–898.

[25] J. Karátson, Characterizing Mesh Independent Quadratic Convergence of Newton's Method for a Class of Elliptic Problems. *J. Math. Anal.* **44** (2012) 1279–1303.

[26] C.T. Kelley, Iterative methods for linear and nonlinear equations. In vol. 16. *SIAM Frontiers in Applied Mathematics* (1996).

[27] 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) 2313–2324.

[28] M.G. Larson and A. Målqvist, An adaptive variational multiscale method for convection-diffusion problems. *Commun. Numer. Methods Engrg.* **25** (2009) 65–79.

[29] M.G. Larson and A. Målqvist, A mixed adaptive variational multiscale method with applications in oil reservoir simulation. *Math. Models Methods Appl. Sci.* **19** (2009) 1017–1042.

[30] A. Målqvist, Multiscale methods for elliptic problems. *Multiscale Model. Simul.* **9** (2011) 1064–1086.

[31] A. Målqvist and D. Peterseim, Localization of Elliptic Multiscale Problems. To appear in *Math. Comput.* (2011). Preprint `arXiv:1110.0692v4`.

[32] Y. Mualem, A New Model for Predicting the Hydraulic Conductivity of Unsaturated Porous Media. *Water Resour. Res.* **12** (1976) 513–522.

[33] J.M. Nordbotten, Adaptive variational multiscale methods for multiphase flow in porous media. *SIAM Multiscale Model. Simul.* **7** (2008) 1455–1473.

[34] D. Peterseim, Robustness of Finite Element Simulations in Densely Packed Random Particle Composites. *Netw. Heterog. Media* **7** (2012) 113–126.

[35] D. Peterseim and S.A. Sauter, Finite Elements for Elliptic Problems with Highly Varying, Non-Periodic Diffusion Matrix. *SIAM Multiscale Model. Simul.* **10** (2012) 665–695.

[36] M. Růžička, Nichtlineare Funktionalanalysis. *Oxford Mathematical Monographs*. Springer-Verlag, Berlin, Heidelberg, New York (2004).