NUMERICAL HOMOGENIZATION OF ELLIPTIC MULTISCALE PROBLEMS BY SUBSPACE DECOMPOSITION

RALF KORNHUBER† AND HARRY YSERENTANT‡

Abstract. Numerical homogenization tries to approximate solutions of elliptic partial differential equations with strongly oscillating coefficients by the solution of localized problems over small subregions. We develop and analyze a rapidly convergent iterative method for numerical homogenization that shares this feature with existing approaches and is modeled after the Schwarz method. The method is highly parallelizable and of lower computational complexity than comparable methods that as ours do not make explicit or implicit use of a scale separation.

Key words. numerical homogenization, localization, subspace correction methods

AMS subject classifications. 65N30, 65N55

1. Introduction. Classical homogenization is an established, powerful technique to approximate partial differential equations with highly oscillatory, periodic coefficient functions by computationally feasible problems with homogenized coefficient functions obtained from auxiliary local problems over a periodicity cell [4, 5, 6, 17]. To overcome the restrictive and often unrealistic periodicity assumption, a variety of different strategies to numerical homogenization has been derived over the last two decades. Using periodic homogenization as a guideline, numerical homogenization aims at a modification of standard finite element discretizations that preserve the accuracy known from smooth coefficients functions in the highly oscillatory case. These modifications are derived from auxiliary local problems. A numerical homogenization method is in this way characterized by a decomposition of the given multiscale problem into a global problem associated with a finite element grid and a number of fully decoupled local subproblems. Examples include the variational multiscale method by Hughes et al. [12], the finite element heterogeneous multiscale method by E and Engquist [8], [2, Section 4], and the multiscale finite element method by Hou and Wu [9, 10], a list that is by far not complete or exhaustive. Error estimates for this kind of methods, like that in [1] or [2] for the finite element heterogeneous multiscale method or that in [11] for the multiscale finite element method, are, however, typically restricted to the case of equations with separated scales, not only due to our lack of understanding but partly also by fundamental reasons.

The variant that has recently been presented by Målqvist and Peterseim [14] constitutes in this respect an important exception. It is founded on a comprehensive convergence theory and utilizes no separation of scales at all. The prize to be paid is a comparably large computational effort. As with other methods of this type, Målqvist and Peterseim calculate approximate solutions that are linked to a finite element grid. The choice of this grid determines the accuracy, which is the same as with smooth coefficient functions. They assign to each vertex of the finite elements a basis function, namely the difference of the corresponding piecewise linear finite element basis function of usual kind and its orthogonal projection to a space of rapidly

†Institut für Mathematik, FU Berlin, 14195 Berlin, Germany (kornhuber@math.fu-berlin.de)

‡Institut für Mathematik, TU Berlin, 10623 Berlin, Germany (yserentant@math.tu-berlin.de)

§The authors are grateful to Joscha Podlesny for providing the numerical computations. This research was supported by Deutsche Forschungsgemeinschaft (DFG) through grant SFB 1114.

oscillating functions. In the basic version of their method, these basis functions possess a global support but decay exponentially from one shell of elements surrounding the assigned vertex to the next. Målqvist and Peterseim are therefore able to replace them by local counterparts without sacrificing the accuracy, at first still on the continuous level and later in discretized form. These modified basis functions can be calculated solving a local fine grid problem. Their support consists of a fixed number of shells of elements surrounding the associated node. The number of these shells increases logarithmically with increasing accuracy, that is, with decreasing element size.

In the approach that we propose and analyze in this work, a similar logarithmic effect shows up. Like usual numerical homogenization methods, our method is based on a decomposition of the original multiscale problem into a comparatively small global problem and a set of decoupled local subproblems. We calculate approximations of the solutions by a rapidly convergent iterative procedure. As Målqvist and Peterseim, we start from a finite element triangulation of the domain under consideration, which is, however, not linked to the accuracy and can therefore be flexibly adapted to other needs, ranging from parallelization issues to the adaption to coefficient inhomogeneities. The assigned finite element space serves only for the stabilization of the iterative method. A new approximation is composed in conjugate gradient like manner of the old approximations and a sum of Ritz projections onto certain subspaces of the solution space. One of these subspaces is the standard piecewise linear finite element space associated with the triangulation. It serves for the global exchange of information. The other subspaces are of very local nature and consist of functions that vanish outside the finite elements surrounding a given vertex. The logarithm comes into play here via the number of iterations. We show that logarithmically many iteration steps suffice to reach a given accuracy. The fine structure of the solution can thus with few local corrections be captured with high accuracy. As that of Målqvist and Peterseim, our approach utilizes no separation of scales at all.

In the first step of our analysis, we attack the continuous original problem directly. The solution space is in this basic version the continuous solution space $H_0^1(\Omega)$ and the local subspaces are the Sobolev spaces $H_0^1(\omega_i)$ over the patches surrounding the vertices of the finite elements. The iterates tend then with a speed that does not deteriorate with decreasing size of the finite elements to the solution of the original problem. This observation is of own interest and underlines the conceptual simplicity and general nature of our approach. In a second stage, the infinite dimensional solution space is replaced by a finite element space of arbitrary order associated with a uniform or nonuniform refinement of the original triangulation.

The resulting numerical method requires the storage and handling of information on the fine grid, but only coarse grid information needs to be exchanged globally. It resembles in this respect common methods for numerical homogenization but is based on an entirely different paradigm, on the approximate, iterative solution of the original problem than on the calculation of a low-dimensional subspace of the fine grid reference space with good approximation properties. It is instructive to compare the computational complexity of the method with that of Målqvist and Peterseim. That we need several iteration sweeps is in general more than counterbalanced by the larger number of local degrees of freedom in the approach of Målqvist and Peterseim, that increases roughly with the second or third power of the number of the involved shells for plane respectively spatial problems, and the high number of basis functions whose supports overlap each other. Both methods are based on similar assumptions, essentially on the existence and stability of a quasi-interpolation operator.

The proposed iterative procedure is an example of an additive subspace correction or additive Schwarz method and is analyzed in this framework. The theory of these methods has in essence been brought to completion in the early 1990s. We refer to [19] and to [23] and the monograph [18] and the references cited therein for more information. A more recent work with similar focus as ours is [16]. In the analysis of such subspace correction methods one usually assumes that the underlying solution space is finite dimensional, unlike the case considered here. As byproduct of our analysis we present therefore a slightly modified version of the theory that covers the infinite dimensional case as well. Sequential versions, modeled after the Gauss-Seidel method, are possible, too, and can be analyzed under the same conditions.

2. The equation and the basic iterative process. The problem considered in this work is a second order differential equation in weak form with homogeneous Dirichlet boundary conditions on a polygonal domain Ω in $d = 2$ or 3 space dimensions. Its solution space is the Sobolev space $H_0^1(\Omega)$ and the associated bilinear form reads

$$a(u, v) \;=\; \int_\Omega \nabla u \cdot A \nabla v \, \mathrm{d}x. \tag{2.1}$$

The matrix A is a function of the spatial variable x with measurable entries and assumed to be symmetric positive definite. We assume that

$$\delta \, |\eta|^2 \leq \; \eta \cdot A(x) \eta \; \leq \; M \, |\eta|^2 \tag{2.2}$$

holds for all $\eta \in \mathbb{R}^d$ and almost all or all $x \in \Omega$, where $|\eta|$ denotes the euclidian norm of η and δ and M are positive constants. This guarantees that the bilinear form (2.1) is an inner product on $H_0^1(\Omega)$ which induces a norm $\| \cdot \|$, the energy norm, that is equivalent to the original norm on this space. The Lax-Milgram theorem states under this condition that the boundary value problem

$$a(u, v) \;=\; f^*(v), \quad v \in H_0^1(\Omega), \tag{2.3}$$

possesses for all bounded linear functionals f^* on $H_0^1(\Omega)$ a unique solution u in $H_0^1(\Omega)$. We emphasize, however, that neither the constants M and δ themselves nor their ratio enter into our estimates for the convergence rate of the proposed method for its iterative solution. If anything, only a local version of condition (2.2) is needed, excluding strong anisotropies and large jumps of the coefficient functions.

We cover the domain Ω with a triangulation $\mathcal{T}$. For simplicity we assume that $\mathcal{T}$ consists of triangles in two and of tetrahedrons in three space dimensions, although the argumentation transfers without essential modifications to other types of elements. We assume that the elements in $\mathcal{T}$ are shape regular but do not require that $\mathcal{T}$ is quasiuniform. Associated with $\mathcal{T}$ is the conforming, piecewise linear finite element subspace $\mathcal{S}$ of $H_0^1(\Omega)$. As said, this space does not serve for the discretization of the boundary value problem but only for the representation of the low frequency parts of the functions in the solution space and for the global transport of information in the iterative process. Let $x_1, x_2, \ldots, x_n$ be the vertices of the elements in $\mathcal{T}$. To each of these vertices, we assign the local patch ω_i, the union of the finite elements surrounding x_i, and the local space $H_0^1(\omega_i)$. To simplify the presentation, let

$$\mathcal{V}_0 = \mathcal{S}, \quad \mathcal{V}_i = H_0^1(\omega_i) \;\text{ for } i = 1, \ldots, n. \tag{2.4}$$

Let $P_i : \mathcal{V} \to \mathcal{V}_i$ be the orthogonal projection from the solution space $\mathcal{V} = H_0^1(\Omega)$ to its subspace $\mathcal{V}_i$ in the sense of the inner product (2.1) on $\mathcal{V}$, defined via the equation

$$a(P_i v, v_i) \;=\; a(v, v_i), \quad v_i \in \mathcal{V}_i. \tag{2.5}$$

Introducing the operator

$$T = P_0 + P_1 + \ldots + P_n \tag{2.6}$$

and fixing a starting value $u^{(0)}$, the approximations of the solution u of the boundary value problem (2.3) are then more or less optimally chosen weighted averages

$$w^{(\ell)} = \sum_{\nu=0}^{\ell} \alpha_{\ell\nu} u^{(\nu)}, \quad \sum_{\nu=0}^{\ell} \alpha_{\ell\nu} = 1, \tag{2.7}$$

of the basic iterates

$$u^{(\nu+1)} = u^{(\nu)} + T(u - u^{(\nu)}). \tag{2.8}$$

The corrections $T(u-u^{(\nu)})$ are composed of a globally defined finite element function and the solutions of continuous local problems on the patches ω_i. Aim is to show that the convergence rate of the resulting iterative or semi-iterative methods is in wide limits independent of the underlying triangulation and that already $\sim \ln(1/\varepsilon)$ iteration steps suffice to reduce the energy norm of the error by the factor $1/\varepsilon$.

3. An analysis of additive subspace correction methods. The convergence analysis of additive subspace correction or additive Schwarz methods as the one under consideration here starts from two abstract assumptions that need to be verified in each particular case. First, one assumes that every function in $v \in \mathcal{V}$ can be decomposed into a sum of functions $v = v_0 + v_1 + \ldots + v_n$ in the subspaces $\mathcal{V}_i$ such that

$$\sum_i \|v_i\|^2 \le K_1 \|v\|^2. \tag{3.1}$$

Secondly, one needs that for every such decomposition

$$\|v\|^2 \le K_2 \sum_i \|v_i\|^2 \tag{3.2}$$

holds. These two assumptions imply the following lemma that is central for the analysis of additive Schwarz methods and represents the basic building block of the theory; see the references given at the end of the introduction.

LEMMA 3.1. *The operator $T : \mathcal{V} \to \mathcal{V}$ from* (2.6) *is symmetric with respect to the inner product on $\mathcal{V}$ given by the expression* (2.1). *For all functions $v \in \mathcal{V}$,*

$$1/K_1\, a(v,v) \le a(Tv,v) \le K_2\, a(v,v). \tag{3.3}$$

Proof. The symmetry of the T results from its definition, that is, the symmetry of the projections P_i. Let $v = v_0 + v_1 + \ldots + v_n$ be a decomposition of the function $v \in \mathcal{V}$ as in the assumption (3.1). Then it follows from the Cauchy-Schwarz inequality

$$a(v,v) = \sum_i a(v_i,v) = \sum_i a(v_i,P_iv) \le \Big(\sum_i \|v_i\|^2\Big)^{1/2} \Big(\sum_i \|P_iv\|^2\Big)^{1/2}$$

and therefore, by assumption (3.1),

$$a(v,v) \le K_1 \sum_i \|P_iv\|^2 = K_1 \sum_i a(P_iv,v) = K_1 a(Tv,v).$$

By assumption (3.2) conversely

$$\sum_i a(P_i v, v) \;\le\; \|\sum_i P_i v\|\|v\| \;\le\; \Big(K_2 \sum_i \|P_i v\|^2\Big)^{1/2} \|v\|$$

and, using once more that P_i is an a-orthogonal projection,

$$\sum_i a(P_i v, v) \;\le\; \Big(K_2 \sum_i a(P_i v, v)\Big)^{1/2} \|v\|,$$

that is, with the definition (2.6) of T, the upper estimate in (3.3). □

In the next step of the analysis one usually expands the elements of the solution space $\mathcal{V}$ in an eigenbasis of the operator T. This is no longer possible in the present context because such an eigenbasis does not need to exist in the infinite dimensional case. Therefore we proceed differently here and first recall the notion of the spectrum of a bounded, symmetric linear operator T mapping a Hilbert space $\mathcal{V}$ into itself. The spectrum $\sigma(T)$ of T is the set of all real λ for which the operator $T - \lambda$ does not possess a bounded inverse, mapping $\mathcal{V}$ back to itself. The spectrum of such an operator is a compact subset of the set of the real numbers.

THEOREM 3.2. *The spectrum of the operator* (2.6) *is a subset of the interval*

$$1/K_1 \;\le\; \lambda \;\le\; K_2. \tag{3.4}$$

The condition number κ of the operator, the ratio of the least upper and the greatest lower bound of its spectrum, is therefore less than or equal to $K_1 K_2$.

Proof. We consider first the case $\lambda < 1/K_1$. The bilinear form

$$\langle u, v\rangle \;=\; a(Tu - \lambda u, v)$$

is for these λ by (3.3) an inner product on the Hilbert space $\mathcal{V}$ that induces a norm which is equivalent to the energy norm. The equation

$$a(Tu - \lambda u, v) \;=\; a(f, v), \quad v \in \mathcal{V},$$

possesses therefore by the Riesz representation theorem a unique solution u that can be estimated by the right hand side $f \in \mathcal{V}$. As $Tu - \lambda u = f$, the operator $T - \lambda$ thus possesses a bounded inverse and λ cannot lie in the spectrum of T. Correspondingly, also the values $\lambda > K_2$ cannot belong to the spectrum. □

To continue, we observe that the difference of the solution u of the boundary value problem (2.3) and the weighted averages (2.7) of the iterates (2.8) can be written as

$$u - w^{(\ell)} = \sum_{\nu=0}^{\ell} \alpha_{\ell\nu} (I - T)^\nu (u - u^{(0)}). \tag{3.5}$$

We need therefore best possible estimates for the norm of operator polynomials $p(T)$. The spectral mapping theorem states that the spectrum of such an operator polynomial consists of the values $p(\lambda)$ with λ in the spectrum of T. Since the norm of a bounded, symmetric linear operator coincides with its spectral radius, therefore

$$\|p(T)\| \;=\; \max\{\, |p(\lambda)| \mid \lambda \in \sigma(T)\,\}, \tag{3.6}$$

as in the finite dimensional case. This allows it to estimate the distance between the solution u and its approximations (2.7) in terms of the condition number of T.

LEMMA 3.3. *For the error between the solution u and its approximations* (2.7)

$$\|u - w^{(\ell)}\| \;\le\; \max_{\lambda\in\sigma(T)} |p(\lambda)|\,\|u - u^{(0)}\| \tag{3.7}$$

holds, where $p(\lambda)$ is the polynomial

$$p(\lambda) \;=\; \sum_{\nu=0}^{\ell} \alpha_{\ell\nu}(1-\lambda)^{\nu} \tag{3.8}$$

of degree ℓ associated with the linear combination (2.7) *of the iterates* (2.8).

From here one can proceed as usual and gets the same estimates in terms of the constants K_1 and K_2 as in the finite dimensional case. Inserting the polynomials $p(\lambda) = (1-\omega\lambda)^{\ell}$ with ω optimally chosen, one obtains the Richardson-type iteration

$$w^{(\ell+1)} \;=\; w^{(\ell)} + \omega\, T(u - w^{(\ell)}) \tag{3.9}$$

with starting value $w^{(0)} = u^{(0)}$, whose convergence rate is

$$q \;=\; \frac{\kappa - 1}{\kappa + 1}. \tag{3.10}$$

The polynomial p of degree ℓ that satisfies the normalization condition $p(0) = 1$ and attains the minimum maximal value on an interval $0 < \alpha \le \lambda \le \beta$ is, up to a linear transformation of the variable and the multiplication by a constant factor, the Chebyshev polynomial of degree ℓ, a fact that is widely used in the analysis of Krylov-space methods [7]. Choosing the averaging coefficients $\alpha_{\ell\nu}$ accordingly, one gets from (3.7) and the bounds (3.4) for the spectrum of T the final estimate for the best attainable convergence rate, realized by the conjugate gradient method.

THEOREM 3.4. *If the coefficients $\alpha_{\ell\nu}$ are optimally chosen,*

$$\|u - w^{(\ell)}\| \;\le\; \frac{2\,q^{\ell}}{1 + q^{2\ell}}\,\|u - u^{(0)}\| \tag{3.11}$$

holds, where the convergence rate

$$q \;=\; \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \tag{3.12}$$

is determined by the condition number $\kappa \le K_1 K_2$ of the operator (2.6).

Although much better than the estimate above for the Richardson-type iteration (3.9), the estimate from Theorem 3.4 is still suboptimal in case of few isolated eigenvalues at the boundary of the spectrum and in comparable situations. This is reflected in the convergence behavior of the conjugate gradient method which chooses by construction the best possible linear combination. Effects of this type in a context similar to ours are described and analyzed in [20] and for another class of elliptic operators, with a small number of negative eigenvalues, in [22].

4. The stability of the subspace decomposition. It remains to prove the estimates (3.1) and (3.2), that is, the stability of the decomposition of the solution space $\mathcal{V} = H_0^1(\Omega)$ into the subspaces (2.4). The upper estimate (3.2) is the easy part.

LEMMA 4.1. *For any decomposition $v = v_0 + v_1 + \ldots + v_n$ of a function $v \in \mathcal{V}$ into functions v_i in the subspaces $\mathcal{V}_i$, the estimate* (3.2) *holds with the constant $K_2 = d+2$:*

$$\|v\|^2 \;\le\; (d+2)\sum_{i=1}^{n} \|v_i\|^2. \tag{4.1}$$

Proof. We restrict the energy norm first to the single finite elements $t \in \mathcal{T}$ and prove a local version of the estimate. Since on every finite element t only $d+2$ of the functions v_i, namely $v_0 \in \mathcal{S}$ and the v_i in the spaces $\mathcal{V}_i = H_0^1(\omega_i)$ assigned to the $d+1$ vertices x_i of the given element, are different from zero, one gets with the triangle- and the Cauchy-Schwarz inequality the estimate

$$\|v\|_t^2 \;\le\; (d+2) \sum_{i=1}^{n} \|v_i\|_t^2.$$

Summation over the elements $t \in \mathcal{T}$ yields (4.1). □

The constant K_2 thus depends only on the space dimension d. The construction of a decomposition of the functions $v \in H_0^1(\Omega)$ into sums $v = v_0 + v_1 + \ldots + v_n$ of functions $v_0 \in \mathcal{S}$ and $v_i \in H_0^1(\omega_i)$, $i = 1, \ldots, n$, for which the lower estimate (3.1) holds is a more difficult task and requires some preparations. Let

$$a(x) \;=\; \max\{\, \eta \cdot A(x)\eta \,|\, |\eta| = 1 \,\}. \tag{4.2}$$

Lemma 4.2. *The function $a : \Omega \to \mathbb{R}$ given by the expression* (4.2) *is measurable. For almost all $x \in \Omega$, $\delta \le a(x) \le M$ holds with the constants δ and M from* (2.2).

Proof. As union of the countably many measurable sets

$$\{\, x \,|\, \eta \cdot A(x)\eta > c \,\}, \quad \eta \in \mathbb{Q}^d \text{ and } |\eta| < 1,$$

the set $\{x \,|\, a(x) > c\}$ is measurable for all real c and the function (4.2) therefore measurable. The bounds for its values are an immediate consequence of (2.2). □

We can therefore assign the weighted L_2-norm on $L_2(\Omega)$ given by the expression

$$\|v\|_0^2 \;=\; \int_\Omega |v(x)|^2\, a(x)\, \mathrm{d}x \tag{4.3}$$

to the energy norm $\|\cdot\|$ on $H^1(\Omega)$ induced by the bilinear form (2.1).

Moreover, we assign to each vertex x_i of the elements in $\mathcal{T}$ the nodal basis function φ_i, the continuous, elementwise linear function that takes the value 1 at x_i and vanishes at all other vertices x_j. The support of φ_i is the patch ω_i, the union of the elements in $\mathcal{T}$ with vertex x_i. The φ_i form a partition of unity on Ω. The key tool for the construction of the desired decompositions of the functions in $H_0^1(\Omega)$ is a locally defined quasi-interpolation operator

$$\Pi : H_0^1(\Omega) \;\to\; \mathcal{S}. \tag{4.4}$$

Such quasi-interpolation operators have a long history. A construction that works for energy norms behaving locally, in a sense explained later, like the H^1-seminorm is

$$\Pi v \;=\; \sum_i \alpha_i \varphi_i, \quad \alpha_i \;=\; \frac{1}{\operatorname{vol}\omega_i} \int_{\omega_i} v\, \mathrm{d}x, \tag{4.5}$$

with the mean values α_i of v over the patches ω_i as nodal values of Πv, where the summation extends here only over the indices i assigned to vertices in the interior of the domain Ω. We decompose the functions $v \in H_0^1(\Omega)$ then into the functions

$$v_0 = \Pi v \in \mathcal{S}, \quad v_i = \varphi_i \cdot (v - \Pi v) \in H_0^1(\omega_i) \;\text{ for } i = 1, \ldots, n. \tag{4.6}$$

LEMMA 4.3. *For all functions* $v \in H^1_0(\Omega)$,

$$\sum_{i=1}^{n} \|\varphi_i\,(v-\Pi v)\|^2 \;\le\; 2\,\|\tau\,(v-\Pi v)\|_0^2 \;+\; 2\,\|v-\Pi v\|^2, \tag{4.7}$$

where the function τ *is constant on the interior of the elements in* $\mathcal{T}$ *and there given by*

$$\tau^2 = \sum_{i=1}^{n} |\nabla\varphi_i|^2. \tag{4.8}$$

Proof. For abbreviation, let $w = v - \Pi v$. As $\nabla(\varphi_i w) = w\nabla\varphi_i + \varphi_i\nabla w$ in weak sense and as the coefficient matrix A is symmetric and positive definite,

$$\nabla(\varphi_i w)\cdot A\nabla(\varphi_i w) \;\le\; 2\,(w\nabla\varphi_i)\cdot A(w\nabla\varphi_i) \;+\; 2\,(\varphi_i\nabla w)\cdot A(\varphi_i\nabla w).$$

By the definition of the weight function (4.2) and because of $\varphi_i^2 \le \varphi_i$, this leads to

$$\nabla(\varphi_i w)\cdot A\nabla(\varphi_i w) \;\le\; 2\,|\nabla\varphi_i|^2\,w^2\,a \;+\; 2\,\varphi_i\nabla w\cdot A\nabla w,$$

at least almost everywhere. Because the φ_i form a partition of unity, integration over the domain Ω and summation over the indices i yield the proposition. □

The function τ can on the elements $t \in \mathcal{T}$ because of their shape regularity be estimated from above and below by the reciprocals of their diameters h_t:

$$\tau \;\sim\; \frac{1}{h_t} \;\text{ on } t \in \mathcal{T}. \tag{4.9}$$

To finalize the proof of the lower estimate (3.1), we assume that for all $v \in H^1_0(\Omega)$

$$\|\Pi v\| \;\le\; c_1\|v\|, \quad \|\tau\,(v-\Pi v)\|_0 \;\le\; c_2\|v\| \tag{4.10}$$

holds. The first condition means that the quasi-interpolation operator (4.4) is stable with respect to the energy norm induced by the bilinear form (2.1). The second condition is an approximation property. Let B_i the ball of minimum radius with center x_i that covers the patch ω_i. It is then not particularly difficult to show that the construction (4.5) satisfies these two assumptions if the local counterpart

$$\delta_i\,|\eta|^2 \le\; \eta\cdot A(x)\eta \;\le\; M_i\,|\eta|^2 \tag{4.11}$$

of condition (2.2) holds for almost all $x \in B_i \cap \Omega$ and the ratios $M_i/\delta_i > 0$ remain uniformly bounded. The proof is based on the Poincaré inequality over these balls, on the assumption that for the balls B_i assigned to the x_i on the boundary of Ω

$$\operatorname{vol} B_i \;\le\; c\,\operatorname{vol} B_i\backslash\Omega \tag{4.12}$$

holds with some constant c independent of i, and on the shape regularity of the finite elements. The condition (4.12) excludes slit domains and means that Ω must satisfy an exterior cone condition. Details can be found in an appendix. A more general construction covering under certain conditions also large jumps of the coefficient functions is presented in [16]. It rests upon weighted Poincaré inequalities [15].

Inserting (4.10) into (4.7), the proof of the stability of the given decomposition of the solution space $H^1_0(\Omega)$ into the finite element space $\mathcal{V}_0 = \mathcal{S}$ and the local Sobolev spaces $\mathcal{V}_i = H^1_0(\omega_i)$ is completed.

LEMMA 4.4. *For all functions* $v \in H_0^1(\Omega)$,

$$\|\Pi v\|^2 + \sum_{i=1}^{n} \|\varphi_i\,(v - \Pi v)\|^2 \;\le\; K_1 \|v\|^2, \tag{4.13}$$

with a constant K_1 *that depends only on the constants* c_1 *and* c_2 *from* (4.10).

From Lemma 4.1 and Lemma 4.4 in conjunction with Theorem 3.2 we obtain our final bounds for the spectrum of the operator (2.6) and can summarize our considerations in the following theorem.

THEOREM 4.5. *The condition number* κ *of the operator* (2.6)*, the ratio of the least upper and the greatest lower bound of its spectrum, is bounded by a constant that depends only on the constants* c_1 *and* c_2 *from* (4.10) *and on the space dimension.*

In other words, if a quasi-interpolation operator (4.4) from $H_0^1(\Omega)$ to the given subspace $\mathcal{S}$ (or any other subspace $\mathcal{S}$, wherever it comes from) of $H_0^1(\Omega)$ exists that is stable with respect to the energy norm and satisfies an approximation property as in (4.10), the iterative method from Sect. 2 converges rapidly. One needs not more than

$$\sim\; \ln(1/\varepsilon) \tag{4.14}$$

iterations to reduce the energy norm of the error by the factor $1/\varepsilon$. The number of iterations does not increase faster than the logarithm of the required accuracy.

5. Sequential versions. Based on the same assumptions, it is also possible to analyze variants of the given iterative method that mimic different variants of the Gauss-Seidel method and exhibit qualitatively the same kind of convergence behavior. The single subspaces $\mathcal{V}_i$ are then not as up to now processed in parallel but sequentially, in arbitrary order. Experience shows that such sequential versions converge in general faster than their additive counterparts considered so far. This holds in particular for symmetrized versions, with symmetric error propagation operators like

$$E \;=\; (I - P_n) \ldots (I - P_1)(I - P_0)(I - P_1) \ldots (I - P_n), \tag{5.1}$$

that can again be accelerated by the conjugate gradient method.

For the sake of completeness, we roughly estimate the speed of convergence of the most simple such sequential version, with the error propagation operator

$$E = (I - P_n) \ldots (I - P_1)(I - P_0), \tag{5.2}$$

along the lines given in [19] or [23]. The precise convergence rate is determined in [21]. The crucial assumption is again the existence of a stable decomposition of any function in the solution space $\mathcal{V}$ into a sum of components in the subspaces $\mathcal{V}_i$ in the sense of condition (3.1). It is complemented by the Cauchy-Schwarz type inequality

$$\sum_{j=0}^{n} \sum_{i=0}^{j} a(u_i, v_j) \;\le\; K_3 \Big(\sum_{i=0}^{n} \|u_i\|^2 \Big)^{1/2} \Big(\sum_{j=0}^{n} \|v_j\|^2 \Big)^{1/2} \tag{5.3}$$

for arbitrarily given functions $u_i \in \mathcal{V}_i$, $v_j \in \mathcal{V}_j$, that is again the easy, uncritical part.

LEMMA 5.1. *For arbitrary functions* $u_i \in \mathcal{V}_i$ *and* $v_j \in \mathcal{V}_j$ *the estimate* (5.3) *holds with a constant* $K_3 \le d + 2$*, that can be bounded in terms of the space dimension* d*.*

Proof. As in the proof of Lemma 4.1 we prove at first a local version of the estimate for a single element $t \in \mathcal{T}$. By the Cauchy-Schwarz inequality,

$$\sum_{j=0}^{n} \sum_{i=0}^{j} a(u_i, v_j)|_t \;\le\; \sum_{j=0}^{n} \sum_{i=0}^{j} \|u_i\|_t \|v_j\|_t \;\le\; \Big(\sum_{i=0}^{n} \|u_i\|_t \Big) \Big(\sum_{j=0}^{n} \|v_j\|_t \Big).$$

Because at most $d+2$ functions u_i respectively v_j are different from zero on t,

$$\sum_{j=0}^{n}\sum_{i=0}^{j} a(u_i,v_j)|_t \;\le\; (d+2)\Big(\sum_{i=0}^{n}\|u_i\|_t^2\Big)^{1/2}\Big(\sum_{j=0}^{n}\|v_j\|_t^2\Big)^{1/2}$$

follows. Summation over all $t\in\mathcal{T}$ and another application of the Cauchy-Schwarz inequality now to the outer sum on the right hand side yield the proposition. □

THEOREM 5.2. *Every cycle of the given iterative method reduces the energy norm of the error at least by the factor $\|E\|$, where, with $C = K_1K_3^2$,*

$$\|E\|^2 \;\le\; 1-\frac{1}{C}. \tag{5.4}$$

Proof. Let $v = v_0+v_1+\ldots+v_n$ be a stable decomposition of the function $v\in\mathcal{V}$ into a sum of functions v_j in the subspace $\mathcal{V}_j$ as in condition (3.1). Let $E_{-1}=I$ and set $E_i=(I-P_i)\ldots(I-P_0)$ for $i=0,1,\ldots,n$. Since E_j maps into the a-orthogonal complement of $\mathcal{V}_j$, $a(E_jv,v_j)=0$ for $j=0,1,\ldots,n$. Therefore

$$\|v\|^2 \;=\; \sum_{j=0}^{n} a((I-E_j)v,v_j) \;=\; \sum_{j=0}^{n}\sum_{i=0}^{j} a(P_iE_{i-1}v,v_j).$$

The Cauchy-Schwarz type inequality (5.3) leads thus to the estimate

$$\|v\|^2 \;\le\; K_3\Big(\sum_{i=0}^{n}\|P_iE_{i-1}v\|^2\Big)^{1/2}\Big(\sum_{j=0}^{n}\|v_j\|^2\Big)^{1/2}.$$

With the stability (3.1) of the given decomposition of v, one obtains the estimate

$$\|v\|^2 \;\le\; C\sum_{i=0}^{n}\|P_iE_{i-1}v\|^2.$$

The summands can be expressed as differences:

$$\|P_iE_{i-1}v\|^2 \;=\; \|E_{i-1}v\|^2-\|E_iv\|^2.$$

Because of $E_{-1}=I$ and $E_n=E$ the estimate thus finally reduces to the estimate

$$\|v\|^2 \;\le\; C\,(\|v\|^2-\|Ev\|^2)$$

that holds for all $v\in\mathcal{V}$ and is therefore equivalent to the proposition. □

6. Discrete variants. The infinite dimensional solution space of the original problem has to be replaced by a discrete counterpart to obtain a computationally feasible method. We start from a potentially very strong, uniform or nonuniform refinement $\mathcal{T}'$ of the triangulation $\mathcal{T}$, bridging the scales and resolving the oscillations of the coefficient functions, and a finite element space $\mathcal{S}'\subseteq H_0^1(\Omega)$ that consists of continuous functions whose restrictions to the elements in $\mathcal{T}'$ are polynomials of a given degree $r\ge 1$. Our aim is to calculate the solution $u\in\mathcal{S}'$ of the equation

$$a(u,v) \;=\; f^*(v),\quad v\in\mathcal{S}', \tag{6.1}$$

approximately by the same kind of iterative methods as before. For this purpose, we need to replace the local solution spaces $\mathcal{V}_i = H_0^1(\omega_i)$ by their discrete counterparts

$$\mathcal{V}_i = \mathcal{S}' \cap H_0^1(\omega_i), \quad i = 1, \ldots, n. \tag{6.2}$$

The coarse grid space $\mathcal{V}_0 = \mathcal{S}$ remains as it is. The resulting numerical method requires the storage and handling of information on the full fine grid but only coarse grid data need to be exchanged globally. It shares this property not with all methods for numerical homogenization, but at least with those that, like that of Målqvist and Peterseim, do not make use of a scale separation or try to exploit statistical effects and that scan the whole information available on the fine grid.

The convergence theory of Sect. 3 applies to this discretized version of our basic method. The upper estimate (4.1) transfers without changes to any decomposition of a function $v \in \mathcal{S}'$ into a sum of a function v_0 in the coarse grid space $\mathcal{S}$ and functions v_i in the local spaces (6.2). The only point that requires special attention is the construction of such a decomposition that is stable in the sense of estimate (3.1). The problem is that the construction from Sect. 4 does not transfer directly to the present case because the product of a function in $\mathcal{S}'$ with a hat functions φ_i is no longer contained in $\mathcal{S}'$. To overcome this problem, we utilize the interpolation operator $\mathcal{I} : C(\bar{\Omega}) \to \mathcal{S}'$ that interpolates at the nodes of usual kind and reproduces the functions in $\mathcal{S}'$. As the operator $\mathcal{I}$ is linear and the φ_i form a partition of unity, we can then decompose the functions $v \in \mathcal{S}'$ into the sum of the functions

$$v_0 = \Pi v \in \mathcal{S}, \quad v_i = \mathcal{I}\left(\varphi_i \left(v - \Pi v\right)\right) \in \mathcal{S}' \cap H_0^1(\omega_i) \ \text{ for } i = 1, \ldots, n. \tag{6.3}$$

The stability of this decomposition in the sense of (3.1) can be easily deduced from the stability of the decomposition of the functions in $H_0^1(\Omega)$ into the sum of the functions (4.6) on condition that that for all functions $v \in \mathcal{S}'$ an estimate

$$\|\mathcal{I}\left(\varphi_i \left(v - \Pi v\right)\right)\| \ \leq \ c_3 \, \|\varphi_i \left(v - \Pi v\right)\| \tag{6.4}$$

holds. The restriction of a function $\varphi_i \left(v - \Pi v\right)$, $v \in \mathcal{S}'$, to an element in $\mathcal{T}'$ is a polynomial of degree $r + 1$. For all $t \in \mathcal{T}'$ and all polynomials p of degree $\leq r + 1$,

$$|\, \mathcal{I} p \,|_{1,\, t} \ \leq \ \theta \, |\, p \,|_{1,\, t}, \tag{6.5}$$

with a constant $\theta > 0$ that depends only on r and the shape regularity of the elements. This proves the estimate (6.4), provided the energy norm can on the elements $t \in \mathcal{T}'$ again be estimated from above and below by the H^1-seminorm $|\cdot|_{1,\, t}$, with constants whose ratio remains uniformly bounded independent of t. We can conclude that there is a constant K_1, now also depending on the constant c_3, such that

$$\|\Pi v\|^2 + \sum_{i=1}^{n} \|\mathcal{I}\left(\varphi_i \left(v - \Pi v\right)\right)\|^2 \ \leq \ K_1 \|v\|^2 \tag{6.6}$$

holds for the decomposition of a function v in the finite element space $\mathcal{S}'$ into the functions (6.3) in the coarse grid space $\mathcal{S}$ and the local spaces (6.2). The rapid convergence of the basic iteration from Sect. 2 thus transfers to the discrete case. The same holds for the Gauss-Seidel type variants from Sect. 5 as the Cauchy-Schwarz type inequality (5.2) is not affected by the transition to the smaller local spaces.

7. Complexity considerations. To get a rough impression of the numerical complexity of the proposed methods in comparison to existing approaches, we consider a simple model problem that illustrates the basic effects. The domain is the two-dimensional unit square that is uniformly refined into squares of edge length H for the coarse and $h \doteq H\varepsilon$ for the fine grid, where the constant ε fixes the lengthscale on which the coefficient functions oscillate, and the local patches are rectangles and squares of edge length between H and $2H$, squares of edge length $2H$ in the interior. The coupling between the gridsizes is motivated by the fact that the second order derivatives of the solutions typically grow like $\sim 1/\varepsilon$ as ε goes to zero. The fine grid discretization error behaves then, in case of linear elements, like $\sim h/\varepsilon$, that is, like $\sim H$ under the given circumstances. As the convergence rate of our methods is independent of the gridsizes, $\sim |\ln H|$ iteration steps are needed to reach an accuracy of order H. Aim is to estimate the work needed by the method of Målqvist and Peterseim for the same data, that is, with the given coupling of the two gridsizes, for the same accuracy, in terms of the cost of one iteration step.

We assume that the cost for the solution of the coarse grid equation and of the local subproblems increases like $\sim N^\beta$, $\beta \geq 1$, with the number N of unknowns and start from the observation that the total cost of a iteration step is dominated by the cost for the solution of the local subproblems. The reason is that the local systems assigned to the interior nodes are because of $1/H < 2H/h$ for $\varepsilon \leq H$ bigger than the global coarse grid system and that the number of all other operations is bounded by the total number of unknowns. The number of local subproblems is in both methods the same, but the number of unknowns in each of these subproblems is in the method of Målqvist and Peterseim by a factor of order $\sim |\ln H|^2$ larger because the edge length of the subdomains increases there logarithmically relative to basic gridsize H. The total effort of the iterative methods is thus at least by a factor of order $\sim |\ln H|^{2\beta-1}$ lower than with the method of Målqvist and Peterseim, the additional cost for assembling the discretization matrix built up from the new basis functions not yet taken into account. Conversely, one gets with same effort an asymptotically much better approximation of the fine grid solution, no matter how meaningful this is in the given case. These effects get even more pronounced in three space dimensions. Such complexity estimates have, of course, to be handled with care and should not be overestimated as a logarithmic factor cancels rapidly with other quantities and the performance of the methods depends on many parameters. Nevertheless they indicate that the iterative approach might be competitive with existing methods.

Even if such considerations could suggest this, we emphasize once more that there is no link between the choice of the coarse grid and the accuracy of our methods. The coarse grid space serves only for the stabilization of the iteration and for the global transport of information across the region. In the basic version of our methods, with the continuous local solution spaces $H_0^1(\omega_i)$, the iterates converge to the exact solution of the multiscale problem, without any error. This is, of course, no longer the case if these local spaces are, as in Sect. 6, replaced by discrete counterparts. In this case, the iterates converge to the solution of the fine grid equation, so that the choice of the fine grid reference space, and only of this space, determines the attainable accuracy. In our example, the fine grid discretization error behaves like $\sim h/\varepsilon$ for linear elements, at least for smooth coefficient functions oscillating on the lengthscale ε and gridsizes h sufficiently small compared to ε, as numerical homogenization requires. If h is halved, asymptotically also the energy norm of the discretization error is halved, independent of the coarse grid. The L_2-error decreases under the given circumstances

even by the factor four. This makes an essential difference in comparison to methods like that of Målqvist and Peterseim, where a coupling between the two gridsizes is mandatory. In methods like that of Målqvist and Peterseim, the choice of the coarse grid determines the low-dimensional subspace of the fine grid reference space in which an approximation is sought and thereby strictly limits the accuracy.

8. First numerical examples. The number iteration steps needed to reduce the error to the size of the fine grid discretization error determines in the end the efficiency of our methods. The crucial question is therefore how fast our iterative methods converge in reality and how robust they are with respect to the variation of the coefficient functions. In the following, we present the results of a few calculations that illustrate their convergence behavior by means of some simple model problems similar to those in the paper [14] of Målqvist and Peterseim. These examples are taken from a somewhat more comprehensive study [13] on the efficiency of different methods for the numerical solution of elliptic multiscale problems.

The domain under consideration is again the two-dimensional unit square of edge length 1 that is subdivided into squares of edge length H and these then further into squares of edge length h. We are working with piecewise linear elements, that is, subdivide each of the resulting squares once more into two triangles. The equation reads

$$-\nabla \cdot (a \nabla u) \;=\; f, \tag{8.1}$$

to be understood in weak form, with zero Dirichlet boundary conditions. The scalar coefficient function a is assumed to be piecewise constant on a 64×64 square grid, with values that are uniformly distributed random numbers in an interval with lower bound $\delta > 0$ and upper bound M. To simplify the computations a bit, we have replaced the local subdomains ω_i by the slightly larger squares of edge length $2H$ centered around the vertices of the coarse grid triangles, with corresponding modification near the boundary of the domain.

It turns out that already the additive version reaches in a few steps a satisfying accuracy and is, with conjugate gradient acceleration, rather robust with respect to the variation of the coefficient functions expressed by the ratio M/δ. The convergence history for four examples, with coarse gridsize $H = 1/32$, fine gridsize $h = 1/512$, right hand side $f = 1$, and ratios $M/\delta = 1, 10, 10^2, 10^4$, and 10^6 is listed in Table 1. Starting values are the coarse grid finite element solutions. The values given in Table 1 are the factors by which the energy norm distance of the iterate to the exact fine grid solution, the iteration error, has decreased with the given iteration step. The convergence rates do not worsen much from the first example, the Laplace equation with the constant coefficient function $a = 1$, to the examples with more extreme ratios M/δ less and less covered by the theory. Let u be the solution of the underlying continuous problem, u_h the exact solution for gridsize h and $u_{h/2}$ the exact solution for the finer gridsize $h/2$. As the function $u_{h/2} - u_h$ is the a-orthogonal projection of $u - u_h$ to the finite element space assigned to the grid of gridsize $h/2$,

$$\|u_{h/2} - u_h\| \;\leq\; \|u - u_h\|. \tag{8.2}$$

In case of all four examples, not more than three iteration steps are needed to reduce the iteration error to a size less than the energy norm distance of the approximations u_h and $u_{h/2}$ of the solution u and thereby as just shown to a size less than the discretization error. In brief, these examples and numerous further calculations with a large variety of different gridsizes and coefficient functions confirm the predictions of the theory and underline the potential of the iterative approach.

step	$M/\delta = 10^0$	$M/\delta = 10^1$	$M/\delta = 10^2$	$M/\delta = 10^4$	$M/\delta = 10^6$
1	0.206	0.458	0.514	0.524	0.545
2	0.296	0.473	0.515	0.522	0.534
3	0.254	0.415	0.396	0.394	0.389
4	0.260	0.333	0.343	0.353	0.367
5	0.320	0.287	0.357	0.366	0.376
6	0.350	0.431	0.538	0.539	0.533
7	0.456	0.389	0.501	0.513	0.498
8	0.381	0.392	0.411	0.405	0.399

TABLE 1
The factors by which the energy norm of the error is reduced from the last step to the given one

Appendix. Some notions and results from spectral theory. For the convenience of the reader, this appendix shortly summarizes the needed facts from the spectral theory of bounded, symmetric linear operators mapping a Hilbert space $\mathcal{H}$ into itself. A standard source in this field is still [3]. We begin with the definition of the (real) resolvent and the spectrum of such operators, that consists in the infinite dimensional case not necessarily only of eigenvalues.

DEFINITION. *The resolvent of a bounded, symmetric operator $T : \mathcal{H} \to \mathcal{H}$ consists of the real numbers λ for which the operator $T - \lambda$ possesses a bounded inverse. The spectrum $\sigma(T)$ of T consists of those real λ that do not belong to the resolvent.*

A first statement on the structure of the spectrum is the following

THEOREM. *The resolvent of the operator T is an open subset of $\mathbb{R}$ and the spectrum a compact subset of the interval with the endpoints $-\|T\|$ and $\|T\|$.*

Proof. Assume that λ_0 belongs to the resolvent. The equation $(T - \lambda)u = f$ is then equivalent to the equation

$$u - (\lambda - \lambda_0)(T - \lambda_0)^{-1}u \;=\; (T - \lambda_0)^{-1}f.$$

If λ is sufficiently close to λ_0, this equation possesses a unique solution that depends continuously on f and can be represented in form of a Neumann series. The resolvent is therefore an open set. A similar fixed point argument, based on the reformulation

$$u - \lambda^{-1}Tu \;=\; -\lambda^{-1}f$$

of the equation $(T - \lambda)u = f$, shows that the λ outside the given interval belong to the resolvent. This proves the theorem. □

The points in the spectrum can be characterized as follows:

THEOREM. *A value λ belongs to the spectrum of the bounded, symmetric linear operator $T : \mathcal{H} \to \mathcal{H}$ if and only if there exists a sequence of elements $u_n \in \mathcal{H}$ with*

$$\lim_{n\to\infty} \|(T - \lambda)u_n\| \;=\; 0, \quad \|u_n\| \;=\; 1,$$

that is, if λ is a so-called approximate eigenvalue.

Proof. Let λ first belong to the resolvent and let R_λ be the bounded inverse of the operator $T-\lambda$. If the vectors $f_n = (T-\lambda)u_n$ tend then to zero as n goes to infinity, the same holds for the vectors $u_n = R_\lambda f_n$. Thus λ cannot be an approximate eigenvalue and the approximate eigenvalues form a part of the spectrum.

Let λ conversely belong to the spectrum. If λ is an eigenvalue of T, nothing has to be shown. If λ is not an eigenvalue, $T-\lambda$ is injective. Furthermore, the range of $T-\lambda$ is a dense subset of $\mathcal{H}$ as can be seen as follows. Let $(u,(T-\lambda)v) = 0$ for all $v \in \mathcal{H}$. Since T is symmetric, then also $((T-\lambda)u, v) = 0$ for all $v \in \mathcal{H}$. This is only possible for $(T-\lambda)u = 0$, that is, for $u = 0$ by the injectivity of $T-\lambda$. Therefore the inverse operator of $T-\lambda$ mapping the range of $T-\lambda$ back to $\mathcal{H}$ cannot be bounded. Otherwise it could be extended to a bounded inverse of $T-\lambda$ and λ would belong to the resolvent. Thus there is a sequence of elements f_n in the range of $T-\lambda$ such that

$$\lim_{n\to\infty} \|f_n\| \;=\; 0, \quad \|(T-\lambda)^{-1} f_n\| \;=\; 1.$$

The vectors $u_n = (T-\lambda)^{-1} f_n$ have then the norm 1 and the norms $\|(T-\lambda)u_n\|$ tend to zero so that λ is indeed an approximate eigenvalue. □

THEOREM. *At least one of the endpoints of the interval $-\|T\| \le \lambda \le \|T\|$ belongs to the spectrum of the bounded, symmetric linear operator $T : \mathcal{H} \to \mathcal{H}$. Therefore*

$$\|T\| \;=\; \max\{\, |\lambda| \mid \lambda \in \sigma(T) \,\}.$$

That is, the operator norm of T coincides with the spectral radius of T.

Proof. Let $\rho = \|T\|$. Then there exists a sequence of elements $u_n \in \mathcal{H}$ such that

$$\|u_n\| = 1, \quad \lim_{n\to\infty} \|Tu_n\| = \rho.$$

Because of the symmetry of T, then

$$\|(T^2-\rho^2)u_n\|^2 \;=\; \|T(Tu_n)\|^2 - 2\rho^2 (Tu_n, Tu_n) \;+\; \rho^4 \|u_n\|^2$$

holds. Since $\|T\| = \rho$ and $\|u_n\| = 1$, this leads to the estimate

$$\|(T^2-\rho^2)u_n\|^2 \;\le\; \rho^4 - \rho^2 \|Tu_n\|^2$$

whose right hand side tends to zero as n goes to infinity. The real number ρ^2 is therefore an approximate eigenvalue of T^2 and thus contained in the spectrum of T^2. That means that the operator

$$T^2 - \rho^2 \;=\; (T-\rho)(T+\rho)$$

does not possess a bounded inverse. This then also holds for one of the operators $T-\rho$ or $T+\rho$. Thus $-\rho$, ρ, or both values are contained in the spectrum of T. □

Next we insert bounded, symmetric operators $T : \mathcal{H} \to \mathcal{H}$ into real polynomials and prove the spectral mapping theorem.

LEMMA. *Let the polynomial $p(\lambda) = \lambda^2 + 2a\lambda + b$ have no real zeroes. The operator*

$$p(T) \;=\; T^2 + 2aT + b$$

possesses then a bounded inverse, independent of the properties of the operator T.

Proof. We first rewrite $p(T)$ in the form

$$p(T) \;=\; (T+a)^2 + \delta, \quad \delta \;=\; b - a^2.$$

Because $p(\lambda)$ has no real zeroes, $\delta > 0$ must hold. The expression

$$\langle u, v \rangle \;=\; (p(T)u, v)$$

defines therefore an inner product that induces a norm which is equivalent to the original norm. By the Riesz representation theorem, thus there exists for every $f \in \mathcal{H}$ a unique $u \in \mathcal{H}$ with $\langle u, v \rangle = (f, v)$ for all $v \in \mathcal{H}$, that is, a unique solution of the equation $p(T)u = f$ whose norm can be bounded by the norm of f. □

THEOREM. *Let $p(\lambda) = a_0 + a_1\lambda + \ldots + a_n\lambda^n$ be a real polynomial and T a symmetric, bounded linear operator. The spectrum of the symmetric, bounded operator*

$$p(T) = a_0 + a_1 T + \ldots + a_n T^n$$

consists then of the values $p(\lambda)$ with λ in the spectrum $\sigma(T)$ of T.

Proof. For constant polynomials $p(\lambda) = a_0$, $p(T)$ is the corresponding multiple of the identity operator whose spectrum consists of a single point, namely $\lambda = a_0$. This proves the theorem for this particular case. We can therefore restrict ourselves to polynomials of degree $n \geq 1$ and can moreover assume that $a_n = 1$.

We first show that any value μ in the spectrum of $p(T)$ is of the form $\mu = p(\lambda)$ with a λ in the spectrum of T. The polynomial $p(\lambda) - \mu$ can over the real numbers be factorized into a product of linear factors $\lambda - \lambda_i$ and of quadratic polynomials $q_i(\lambda)$ without real zeroes, which turns into a factorization of the operator $p(T) - \mu$. This operator possesses by the previous lemma a bounded inverse if the operators $T - \lambda_i$ possess bounded inverses, that is, if none of the solutions λ_i of the equation $p(\lambda) = \mu$ is contained in the spectrum of T. Any μ that belongs to the spectrum of $p(T)$ must therefore be of the form $\mu = p(\lambda)$ with some λ in the spectrum of T.

Let λ conversely be contained in the spectrum of T and let $\mu = p(\lambda)$. Then there exists a sequence of elements $u_k \in \mathcal{H}$ of norm 1 for which the norms of the $(T - \lambda)u_k$ tend to zero as k goes to infinity. As $p(\lambda) - \mu = 0$, $p(\xi) - \mu = q(\xi)(\xi - \lambda)$ with a polynomial $q(\xi)$ of degree $n - 1$. The sequence of the vectors

$$(p(T) - \mu)u_k \;=\; q(T)(T - \lambda)u_k$$

tends therefore to zero as well. That is, μ is an approximate eigenvalue of $p(T)$ and thus belongs to the spectrum of $p(T)$. □

Hence we can conclude that for any real polynomial $p(\lambda)$ the norm of the operator polynomial $p(T)$ is determined by the values $p(\lambda)$, $\lambda \in \sigma(T)$, and is given by

$$\|p(T)\| \;=\; \max\{\, |p(\lambda)| \mid \lambda \in \sigma(T) \,\},$$

regardless of the existence of a complete set of eigenvectors.

Appendix. The quasi-interpolation operator. The key to our proof is the existence of a quasi-interpolation operator (4.4) that satisfies the two assumptions from (4.10). We prove in this appendix that the operator given by (4.5) falls into this category, provided the energy norm induced by the bilinear form (2.1) behaves locally, in the sense of condition (4.11), like the H^1-seminorm.

THEOREM. *The quasi-interpolation operator given by* (4.5) *satisfies the conditions from* (4.10)*. The constants depend only on an upper bound for the ratio M_i/δ_i of the constants in the assumption* (4.11)*, on the degeneration of the finite elements, and on the geometry of the domain Ω via the local condition* (4.12) *to its boundary.*

Proof. It suffices to prove the estimates (4.10) for continuously differentiable functions $v : \mathbb{R}^n \to \mathbb{R}$ that vanish outside a compact subset of Ω. The rest then follows by the density of these in $H_0^1(\Omega)$. We use the notation $a \lesssim b$, meaning that a can be estimated by b up to a constant that depends at most on the shape regularity of the finite elements and the constant from (4.12). As the φ_i form a partition of unity,

$$v - \Pi v \;=\; \sum_{i \in \mathcal{N}} \varphi_i \, (v - \alpha_i) \;+\; \sum_{i \notin \mathcal{N}} \varphi_i \, v,$$

where $\mathcal{N}$ is the set of the indices of the vertices in the interior of Ω. As on a given element only the $d+1$ functions φ_i assigned to its vertices are different from zero,

$$\|v - \Pi v\|^2 \;\lesssim\; \sum_{i \in \mathcal{N}} \|\varphi_i \, (v - \alpha_i)\|^2 \;+\; \sum_{i \notin \mathcal{N}} \|\varphi_i v\|^2.$$

Let h_i be the radius of the ball B_i. As $|\nabla \varphi_i| \lesssim h_i^{-1}$ by the shape regularity of the elements and $0 \le \varphi_i \le 1$, the terms in the first sum can by (4.11) be estimated as

$$\|\varphi_i \, (v - \alpha_i)\| \;\lesssim\; M_i h_i^{-1} \|v - \alpha_i\|_{L_2(\omega_i)} + M_i \, \|\nabla v\|_{L_2(\omega_i)}.$$

The constant α_i is the mean value of the function v over the patch ω_i. Let α_i' be the mean value of v over the ball B_i. Because

$$\|v - \alpha_i\|_{L_2(\omega_i)} \;\le\; \|v - \alpha_i'\|_{L_2(\omega_i)} \;\le\; \|v - \alpha_i'\|_{L_2(B_i)},$$

the Poincaré inequality for balls leads therefore to the estimate

$$\|v - \alpha_i\|_{L_2(\omega_i)} \;\lesssim\; h_i \, \|\nabla v\|_{L_2(B_i)}.$$

For the terms associated with the inner vertices, thus finally

$$\|\varphi_i \, (v - \alpha_i)\| \;\lesssim\; M_i \, \|\nabla v\|_{L_2(B_i)} \;\le\; M_i \delta_i^{-1} \|v\|_{B_i \cap \Omega}.$$

For a boundary term correspondingly

$$\|\varphi_i v\| \;\lesssim\; M_i h_i^{-1} \|v\|_{L_2(\omega_i)} + M_i \, \|\nabla v\|_{L_2(\omega_i)}$$

holds. The mean value of the given functions v over the part of the assigned ball B_i outside of Ω is zero. The L_2-distance over B_i of v to this mean value can again be estimated by means of the Poincaré inequality for balls. By (4.12), for this reason

$$\|v\|_{L_2(B_i)} \;\lesssim\; h_i \, \|\nabla v\|_{L_2(B_i)},$$

with a constant depending on that from (4.12). For the boundary terms therefore

$$\|\varphi_i v\| \;\lesssim\; M_i \, \|\nabla v\|_{L_2(B_i)} \;\le\; M_i \delta_i^{-1} \|v\|_{B_i \cap \Omega}.$$

As $M_i / \delta_i \le K$ for some constant K and as the balls B_i form, because of the shape regularity of the finite elements, a locally finite covering of Ω, we get

$$\|v - \Pi v\| \;\lesssim\; K \, \|v\|,$$

which implies the stability of Π. Using that for all square integrable functions w

$$\|\tau \varphi_i w\|_0 \;\lesssim\; M_i h_i^{-1} \|w\|_{L_2(\omega_i)}$$

holds, the approximation property follows by the same arguments. □

REFERENCES

[1] A. Abdulle, *A priori and a posteriori error analysis for numerical homogenization: a unified framework*, Series in Contemporary Applied Mathematics, 16 (2011), pp. 280–305.

[2] A. Abdulle, W. E, B. Engquist, and E. Vanden-Eijnden, *The hetereogeneous multiscale method*, Acta Numerica, 21 (2012), pp. 1–87.

[3] N. I. Achieser and I. M. Glasmann, *Theorie der linearen Operatoren im Hilbert-Raum*, Harri Deutsch, Thun, 1981. German translation of the Russian original.

[4] G. Allaire, *Homogenization and two-scale convergence*, SIAM J. Math. Anal., 23 (1992), pp. 1482–1512.

[5] N. Bakhvalov and G. Panasenko, *Homogenization: Averaging Processes in Periodic Media*, vol. 36 of Mathematics and its applications, Kluwer, Dordrecht, 1990.

[6] D. Cioranescu and P. Donato, *An introduction to Homogenization*, vol. 17 of Oxford Lecture Series in Mathematics and Applications, Oxord University Press, Oxford, 1999.

[7] P. Deuflhard and A. Hohmann, *Numerical Analysis in Modern Scientific Computing: An Introduction*, vol. 43 of Texts in Applied Mathematics, Springer, Berlin Heidelberg, 2003.

[8] W. E and B. Engquist, *The hetereogeneous multiscale methods*, Comm. Math. Sci., 1 (2003), pp. 87–132.

[9] Y. Efendiev and T. Y. Hou, *Multiscale Finite Element Methods: Theory and Applications*, vol. 4 of Surveys and Tutorials in the Applied Mathematical Sciences, Springer, New York, 2009.

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

[11] T. Y. Hou, X.-H. Wu, and Z. Cai, *Convergence of a multiscale finite element method for elliptic problems with rapidly oscillating coefficients*, Math. Comp., 68 (1999), pp. 913–943.

[12] T. J. R. Hughes, G. R. Feijó, L. M. Mazzei, and J.-B. Quincy, *The variational multiscale method - a paradigm for computational mechanics*, Comput. Methods Appl. Mech. Engrg., 166 (1998), pp. 3–24.

[13] R. Kornhuber, J. Podlesny, and H. Yserentant, *A comparative study of some direct and iterative methods for numerical homogenization.* In preparation.

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

[15] C. Pechstein and R. Scheichl, *Weighthed Poincaré inequalities*, IMA J. Numer. Anal., 33 (2013), pp. 652–686.

[16] R. Scheichl, P. Vassilevski, and L. Zikatanov, *Multilevel methods for elliptic problems with highly varying coefficients on nonaligned coarse grids*, SIAM J. Numer. Anal., 50 (2012), pp. 1675–1694.

[17] L. Tartar, *The General Theory of Homogenization. A Personalized Introduction*, vol. 7 of Lecture Notes of the Unione Matematica Italiana, Springer, Berlin, 2009.

[18] A. Toselli and O. Widlund, *Domain Decomposition Methods - Algorithms and Theory*, vol. 34 of Springer Series in Computational Mathematics, Springer, Berlin Heidelberg, 2005.

[19] J. Xu, *Iterative methods by space decomposition and subspace correction*, SIAM Review, 34 (1992), pp. 581–613.

[20] J. Xu and Y. Zhu, *Uniform convergent multigrid methods for elliptic problems with strongly discontinuous coefficients*, Math. Models Methods Appl. Sci., 18 (2008), pp. 77–105.

[21] J. Xu and L. Zikatanov, *The method of alternating projections and the method of subspace corrections in Hilbert space*, J. Amer. Math. Soc., 15 (2002), pp. 573–597.

[22] H. Yserentant, *Preconditioning indefinite discretization matrices*, Numer. Math., 54 (1988), pp. 719–734.

[23] ———, *Old and new convergence proofs for multigrid methods*, Acta Numerica, 2 (1993), pp. 285–326.