Calcolo (2021) 58:31
https://doi.org/10.1007/s10092-021-00413-w

Approximating inverse FEM matrices on non-uniform meshes with $\mathcal{H}$-matrices

Niklas Angleitner[1] · Markus Faustmann[1] · Jens Markus Melenk[1]

Received: 9 May 2020 / Revised: 14 January 2021 / Accepted: 12 April 2021 /
Published online: 30 June 2021

Abstract
We consider the approximation of the inverse of the finite element stiffness matrix in the data sparse $\mathcal{H}$-matrix format. For a large class of shape regular but possibly non-uniform meshes including algebraically graded meshes, we prove that the inverse of the stiffness matrix can be approximated in the $\mathcal{H}$-matrix format at an exponential rate in the block rank. Since the storage complexity of the hierarchical matrix is logarithmic-linear and only grows linearly in the block-rank, we obtain an efficient approximation that can be used, e.g., as an approximate direct solver or preconditioner for iterative solvers.

Keywords FEM · $\mathcal{H}$-matrices · Approximability · Non-uniform meshes

Mathematics Subject Classification Primary: 65F50 · Secondary: 65F30, 65N30

1 Introduction

Discretizations of elliptic partial differential equations on a domain $\Omega \subseteq \mathbb{R}^d$ using the classical finite element method (FEM) usually produce sparse linear systems of equations $\boldsymbol{Ax} = \boldsymbol{b}$ with storage requirements linear in the number of unknowns and linear complexity for the matrix-vector multiplication. However, the direct solution of these systems is computationally more expensive. Therefore, iterative solution methods (e.g., Krylov space methods) are popular in applications, since

✉ Niklas Angleitner
niklas.angleitner@tuwien.ac.at

Markus Faustmann
markus.faustmann@tuwien.ac.at

Jens Markus Melenk
melenk@tuwien.ac.at

[1] Technische Universität Wien, Institute of Analysis and Scientific Computing (Inst. E 101), Wiedner Hauptstrasse 8-10, A–1040 Wien, Austria

they only need matrix-vector multiplications, which can be done in linear complexity. A drawback of these methods is that convergence can be slow for matrices with large condition numbers unless a suitable preconditioner is employed. These preconditioners have to be taylored to the problem at hand making black box preconditioners that are based on (approximate) direct solvers particularly interesting. Moreover, if one is interested in solving the same problem with (many) different right-hand sides, a direct solver may be computationally advantageous.

Hierarchical matrices ($\mathcal{H}$-matrices), introduced in [20] and extensively studied in the monograph [21], provide a different solution approach to this problem that does not suffer from the drawbacks of classical direct and iterative methods. $\mathcal{H}$-matrices are blockwise low-rank matrices. For suitable block structures and block ranks, storing an $\mathcal{H}$-matrix is of logarithmic-linear complexity. Approximating a given matrix in the $\mathcal{H}$-matrix format thus effects a compression. A main difference to other compression methods such as multipole expansions, [18, 23], or wavelet methods, [24–26], is that the $\mathcal{H}$-matrix format allows for an approximate arithmetic. It is possible to add and multiply as well as compute inverses and LU-decompositions efficiently in the format, [14, 19, 21]. Therefore, using an $\mathcal{H}$-matrix approximation to the inverse $\boldsymbol{A}^{-1}$ gives an approximate direct solution method of logarithmic-linear complexity that can be applied efficiently to multiple right-hand sides. Moreover, an LU-decomposition in the $\mathcal{H}$-matrix format can be used as a black-box preconditioner in iterative solvers, [3, 15, 17, 22]. Nonetheless, we mention that the accuracy in terms of the maximal blockwise rank of the computed approximations to $\boldsymbol{A}^{-1}$ (or the LU-decomposition) using $\mathcal{H}$-matrix arithmetic is not fully understood yet.

In order to explain the numerical success of these approximations, first observed in [19], several works in the literature provide existence results of approximations to the inverse matrices in the $\mathcal{H}$-matrix format. See, e.g., [2, 5, 7, 11] for the inverses of FEM matrices and [12, 13] for the inverses of BEM matrices. These analyses are restricted to the case of (quasi)uniform meshes, i.e., all mesh elements have comparable size. In a typical FEM scenario, however, locally refined meshes are employed with mesh elements varying greatly in size in order to account for effects such as locally reduced regularity of the solution. A classical example are graded meshes for the solution of elliptic problems in corner domains, [6].

In this article, we generalize the results of [11] for quasiuniform meshes to meshes of so called *locally bounded cardinality* (cf. Definition 2.4), which includes both uniform meshes and algebraically graded meshes. Our main result states that the inverses of FEM matrices for such meshes can be approximated by hierarchical matrices such that the error converges exponentially in the $\mathcal{H}$-matrix block rank r. Given a clustering strategy suitable for non-uniform grids, cf. [16], the storage complexity of the $\mathcal{H}$-matrix approximant is of logarithmic-linear complexity $\mathcal{O}(rN \ln N)$. Moreover, we develop an abstract framework that allows for more general FEM basis functions that do not need to have local supports. In fact, locality is necessary only for a set of *dual functions*, which is a substantially weaker assumption. Finally, we streamline some of the arguments made in [11]. While not repeated in this article, we mention that the (mostly algebraic) techniques of [11, Section 5] can be employed in exactly the same way to derive exponentially convergent approximate LU-decompositions in the $\mathcal{H}$-matrix format.

The present paper is structured as follows: In Sect. 2 we introduce all necessary definitions and concepts and state our main result, Theorem 2.15. Section 3 is dedicated to the proof of the main result. The main technical contribution is the discrete Caccioppoli-type estimate presented in Lemma 3.27, which is of independent interest. For a certain class of functions, it allows us to bound the H^1-seminorm on a given subdomain by the L^2-norm on a slightly larger subdomain. Finally, Sect. 4 provides numerical examples that illustrate our main result.

Concerning notation: We write "$a \lesssim b$", if there exists a constant $C > 0$ such that "$a \leq Cb$". The constant might depend on the space dimension d, the domain Ω, the coefficients of the PDE, the shape regularity constant of the mesh, and the polynomial degree of the discrete spline space, but it is *in*dependent of all critical parameters such as the mesh width. We write $a \eqsim b$, if there hold both $a \lesssim b$ and $a \gtrsim b$. Matrices and vectors in linear systems of equations are expressed in boldface letters, e.g., $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ and $\boldsymbol{f} \in \mathbb{R}^N$. For all $x \in \mathbb{R}^d$ and $\varepsilon > 0$, we write $\mathrm{Ball}_2(x, r) := \{y \in \mathbb{R}^d \mid \|y - x\|_2 < \varepsilon\}$ for the Euclidean ball of radius r centered at x. The norm of the sequence spaces ℓ^1 and ℓ^2 is denoted by $\|\cdot\|_1$ and $\|\cdot\|_2$. For $k \geq 0$, $q \in [1, \infty]$ and domains $\Omega \subseteq \mathbb{R}^d$, we denote the Sobolev space by $W^{k,q}(\Omega)$. For a given mesh $\mathcal{T}$, we denote by $W^{k,q}_{\mathrm{pw}}(\mathcal{T})$ the broken Sobolev space consisting of elementwise functions from $W^{k,q}$. For all $v \in W^{k,q}_{\mathrm{pw}}(\mathcal{T})$ and $\mathcal{B} \subseteq \mathcal{T}$, we set $|v|_{W^{k,q}(\mathcal{B})} := (\sum_{T\in\mathcal{B}} |v|^q_{W^{k,q}(T)})^{1/q}$ and $|v|_{W^{k,\infty}(\mathcal{B})} := \max_{T\in\mathcal{B}} |v|_{W^{k,\infty}(T)}$. Similarly, $C^0_{\mathrm{pw}}(\mathcal{T})$ denotes the space of piecewise continuous functions. For all $v \in L^2(\Omega)$ and $\mathcal{B} \subseteq \mathcal{T}$, the restriction of v to $\bigcup \mathcal{B} \subseteq \mathbb{R}^d$ is abbreviated as $v|_{\mathcal{B}} := v|_{\bigcup \mathcal{B}}$. Finally, it will facilitate notation on numerous occasions to define the *(discrete) support* of a function $v \in L^2(\Omega)$ on a mesh $\mathcal{T}$ by $\mathrm{supp}_{\mathcal{T}}(v) := \{T \in \mathcal{T} \mid v|_T \not\equiv 0\}$. In particular, we have $\mathrm{supp}_{\mathcal{T}}(v) \subseteq \mathcal{T}$ and $\bigcup \mathrm{supp}_{\mathcal{T}}(v) \subseteq \mathbb{R}^d$, which slightly differs from the usual definition of a support, namely, $\mathrm{supp}(v) := \{x \in \Omega \mid v(x) \neq 0\} \subseteq \mathbb{R}^d$.

2 Main results

2.1 The model problem

We investigate the following *model problem*: Let $d \geq 1$ and $\Omega \subseteq \mathbb{R}^d$ be a bounded polyhedral Lipschitz domain. Furthermore, let $a_1 \in L^\infty(\Omega, \mathbb{R}^{d\times d})$, $a_2 \in L^\infty(\Omega, \mathbb{R}^d)$ and $a_3 \in L^\infty(\Omega, \mathbb{R})$ be given coefficient functions and $f \in L^2(\Omega)$ be a given right-hand side. We seek a weak solution $u \in H^1_0(\Omega)$ to the following equations:

$$\begin{aligned} -\mathrm{div}(a_1 \cdot \nabla u) + a_2 \cdot \nabla u + a_3 u &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega. \end{aligned}$$

In the present work, we restrict ourselves to homogeneous Dirichlet conditions. For the treatment of Neumann and Robin boundary conditions, the same arguments as in [11] can be employed.

We assume that a_1 is coercive in the sense $\langle a_1(x)y, y\rangle \geq \alpha_1 \|y\|_2^2$ for all $x \in \Omega$, $y \in \mathbb{R}^d$ and some constant $\alpha_1 > \sigma^2_{\mathrm{Pcr}}(\|a_2\|_{L^\infty(\Omega)} + \|a_3\|_{L^\infty(\Omega)}) \geq 0$. Here, $\sigma_{\mathrm{Pcr}} > 0$ denotes the constant in the Poincaré inequality $\|\cdot\|_{H^1(\Omega)} \leq \sigma_{\mathrm{Pcr}} |\cdot|_{H^1(\Omega)}$ on $H^1_0(\Omega)$.

Definition 2.1 We introduce the bilinear form:

$$\forall u, v \in H_0^1(\Omega) : \qquad a(u,v) := \langle a_1 \nabla u, \nabla v \rangle_{L^2(\Omega)} + \langle a_2 \cdot \nabla u, v \rangle_{L^2(\Omega)} + \langle a_3 u, v \rangle_{L^2(\Omega)}.$$

The weak formulation of the *model problem* reads as follows: Find $u \in H_0^1(\Omega)$ such that

$$\forall v \in H_0^1(\Omega) : \qquad a(u,v) = \langle f, v \rangle_{L^2(\Omega)}.$$

The assumptions on the PDE coefficients imply that the bilinear form $a(\cdot,\cdot)$ is continuous and coercive, cf. Lemma 3.7. In particular, the well-known Lax-Milgram Lemma yields the existence of a unique solution $u \in H_0^1(\Omega)$.

2.2 The mesh

Throughout the text, we consider regular, affine meshes in the following sense:

Definition 2.2 *(Mesh)* A finite set $\mathcal{T} \subseteq \mathrm{Pow}(\Omega)$ is a *mesh* if there exists an open simplex $\hat{T} \subseteq \mathbb{R}^d$ (the *reference element*) such that every *element* $T \in \mathcal{T}$ is of the form $T = F_T(\hat{T})$, where $F_T : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ is an affine diffeomorphism. Furthermore, the elements must be pairwise disjoint, i.e., $|T \cap S| = 0$ for all $T \neq S \in \mathcal{T}$, and constitute a partition of Ω, i.e., $\bigcup_{T \in \mathcal{T}} \overline{T} = \overline{\Omega}$. Finally, a mesh must be regular in the sense of [9], i.e., it does not contain any hanging nodes.

We call a collection of mesh elements $\mathcal{B} \subseteq \mathcal{T}$ a *cluster*. In the literature on hierarchical matrices, the word *cluster* is typically reserved for collections of vector/matrix indices $I \subseteq \{1, \dots, N\}$. In the present work, however, we deal with collections of mesh elements $\mathcal{B} \subseteq \mathcal{T}$ much more frequently. We also note that both concepts are intimately linked via Definition 2.8.

For every subset $B \subseteq \mathbb{R}^d$, we call the set of neighboring mesh elements

$$\mathcal{T}(B) := \{T \in \mathcal{T} \,|\, \overline{T} \cap \overline{B} \neq \emptyset\} \subseteq \mathcal{T} \tag{2.1}$$

the *patch* of B. Similarly, for every cluster $\mathcal{B} \subseteq \mathcal{T}$, we set $\mathcal{T}(\mathcal{B}) := \bigcup_{B \in \mathcal{B}} \mathcal{T}(B) \subseteq \mathcal{T}$.

To measure the size of an element $T \in \mathcal{T}$, we introduce the local *mesh width* $h_T := \sup_{x,y \in T} \|y - x\|_2$. The corresponding aggregate mesh widths for a cluster $\mathcal{B} \subseteq \mathcal{T}$ read $h_{\mathcal{B}} := h_{\max,\mathcal{B}} := \max_{T \in \mathcal{B}} h_T$ and $h_{\min,\mathcal{B}} := \min_{T \in \mathcal{B}} h_T$.

Finally, for every $T \in \mathcal{T}$, we denote the center of the largest inscribable ball by $x_T \in T$ (the *incenter*). We assume that $\mathcal{T}$ is part of a *shape-regular* family of meshes, i.e., there exists a constant $\sigma_{\mathrm{shp}} \geq 1$ such that

$$\forall T \in \mathcal{T} : \qquad \mathrm{Ball}_2(x_T, \sigma_{\mathrm{shp}}^{-1} h_T) \subseteq T \subseteq \bigcup \mathcal{T}(T) \subseteq \mathrm{Ball}_2(x_T, \sigma_{\mathrm{shp}} h_T).$$

Definition 2.3 *(Mesh metric)* The *mesh metric* is given by

$$\forall T, S \in \mathcal{T}: \qquad \operatorname{dist}_{\mathcal{T}}(T,S) := \|x_S - x_T\|_2.$$

For all clusters $\mathcal{A}, \mathcal{B} \subseteq \mathcal{T}$, we denote the corresponding diameters and distances by

$$\operatorname{diam}_{\mathcal{T}}(\mathcal{A}) := \max_{A_1, A_2 \in \mathcal{A}} \operatorname{dist}_{\mathcal{T}}(A_1, A_2), \qquad \operatorname{dist}_{\mathcal{T}}(\mathcal{A}, \mathcal{B}) := \min_{\substack{A \in \mathcal{A}, \\ B \in \mathcal{B}}} \operatorname{dist}_{\mathcal{T}}(A, B).$$

If $\mathcal{A}$ or $\mathcal{B}$ contains only one element, e.g., $\mathcal{A} = \{T\}$, we drop the enclosing braces and simply write $\operatorname{dist}_{\mathcal{T}}(T, \mathcal{B}) := \operatorname{dist}_{\mathcal{T}}(\{T\}, \mathcal{B})$. Furthermore, $\operatorname{diam}_{\mathcal{T}}(T) := \operatorname{diam}_{\mathcal{T}}(\{T\}) = 0$ by definition of the cluster diameter.

We refer to Lemma 3.15 for some basic properties of the mesh metric.

Compared to [11], we consider a more general class of meshes. Here, the crucial property is the so called *locally bounded cardinality* defined in the following Definition 2.4. In Sect. 3.2, we provide examples of meshes both with (uniform and algebraically graded meshes) and without said property (exponentially graded meshes).

Definition 2.4 *(Locally bounded cardinality)* A mesh $\mathcal{T} \subseteq \operatorname{Pow}(\Omega)$ has *locally bounded cardinality*, if there exists a constant $\sigma_{\mathrm{card}} \geq 1$ such that

$$h_{\mathcal{T}}^{\sigma_{\mathrm{card}}} \lesssim h_{\min,\mathcal{T}}, \qquad \forall \mathcal{B} \subseteq \mathcal{T}: \quad \#\mathcal{B} \lesssim \left(1 + \frac{\operatorname{diam}_{\mathcal{T}}(\mathcal{B})}{h_{\mathcal{B}}}\right)^{d\sigma_{\mathrm{card}}}.$$

2.3 The basis and dual functions

Definition 2.5 *(Spline spaces)* Let $k \geq 0$ and $p \geq 0$. We introduce the finite-dimensional *spline spaces*

$$\mathbb{S}^{p,k}(\mathcal{T}) := \{v \in H^k(\Omega) \mid \forall T \in \mathcal{T}: \ v \circ F_T \in \mathbb{P}^p(\hat{T})\},$$
$$\mathbb{S}_0^{p,k}(\mathcal{T}) := \mathbb{S}^{p,k}(\mathcal{T}) \cap H_0^1(\Omega),$$

where $\mathbb{P}^p(\hat{T}) := \operatorname{span}\{\hat{T} \ni x \mapsto x^q \mid \|q\|_1 \leq p\}$ denotes the usual space of polynomials of (total) degree p on the reference element.

The following definition introduces the bases of $\mathbb{S}_0^{p,1}(\mathcal{T})$ that we consider:

Definition 2.6 *(Basis with local dual functions)* Let $p \geq 1$, $N := \dim \mathbb{S}_0^{p,1}(\mathcal{T})$ and $\{\varphi_1, \ldots, \varphi_N\} \subseteq \mathbb{S}_0^{p,1}(\mathcal{T})$ be a basis. We say that the basis *allows for a system of local dual functions*, if there exist functions $\{\lambda_1, \ldots, \lambda_N\} \subseteq L^2(\Omega)$ with the following properties:

1. Duality: For all $n, m \in \{1, \ldots, N\}$, there holds $\langle \varphi_n, \lambda_m \rangle_{L^2(\Omega)} = \delta_{nm}$.

2. Stability: For all $\boldsymbol{x} \in \mathbb{R}^N$, there holds the bound $\| \sum_{m=1}^N \boldsymbol{x}_m \lambda_m \|_{L^2(\Omega)} \lesssim h_{\min,\mathcal{T}}^{-d/2} \|\boldsymbol{x}\|_2$. The implied constant may only depend on d, p, and the shape regularity of the mesh $\mathcal{T}$.
3. Locality: For every $n \in \{1, \dots, N\}$, there exists a characteristic element $T_n \in \mathcal{T}$ such that $\operatorname{supp}_{\mathcal{T}}(\lambda_n) \subseteq \mathcal{T}(T_n)$. For every $T \in \mathcal{T}$, there holds the uniform bound $\#\{n \mid T_n = T\} \lesssim 1$.

Remark 2.7 Note that we *do not* assume *local* basis functions φ_n, i.e., $\operatorname{supp}_{\mathcal{T}}(\varphi_n) = \mathcal{T}$ is allowed. Rather, locality is imposed only on the dual functions. In the *finite element* framework described in Sect. 3.3, this distinction might seem unnecessary, as both the basis functions and the dual functions are indeed local. In the somewhat similar setting of *radial basis functions* (see, e.g., [27]), however, the distinction becomes crucial. There, the basis functions have global supports by nature and locality can only be imposed on the dual functions. Here, our goal is to formulate the more general framework such that we can apply some results in our upcoming work on $\mathcal{H}$-matrices and radial basis functions (cf. [1]) as well.

The fundamental idea of the present work is to derive properties of matrices from properties of function spaces. Naturally, one has to think about the connection between abstract matrix indices $n \in \{1, \dots, N\}$ and corresponding subdomains of Ω, which is captured in the following definition.

Definition 2.8 *(Index patches)* We define the *index patches*

$$\forall I \subseteq \{1, \dots, N\} : \qquad \mathcal{T}(I) := \bigcup_{n \in I} \operatorname{supp}_{\mathcal{T}}(\lambda_n) \subseteq \mathcal{T}.$$

Recall from Sect. 2.2 that $\mathcal{T}(B) \subseteq \mathcal{T}$ is the patch of a subdomain $B \subseteq \mathbb{R}^d$ and that $\mathcal{T}(\mathcal{B}) \subseteq \mathcal{T}$ is the patch of a cluster $\mathcal{B} \subseteq \mathcal{T}$. We also have patches $\mathcal{T}(I) \subseteq \mathcal{T}$ for collections of matrix indices $I \subseteq \{1, \dots, N\}$. Since all three types of patches follow a common idea, we chose the similarity in notation on purpose.

2.4 The system matrix

Let $\mathcal{T} \subseteq \operatorname{Pow}(\Omega)$ be a mesh and $p \geq 1$ a fixed polynomial degree. Let $\mathbb{S}_0^{p,1}(\mathcal{T}) \subseteq H_0^1(\Omega)$ be the corresponding spline space. We discretize the model problem from Sect. 2.1 by means of the spline space and get the following *discrete model problem*: For given $f \in L^2(\Omega)$, find $u \in \mathbb{S}_0^{p,1}(\mathcal{T})$ such that

$$\forall v \in \mathbb{S}_0^{p,1}(\mathcal{T}) : \qquad a(u, v) = \langle f, v \rangle_{L^2(\Omega)}.$$

Again, existence and uniqueness of a solution $u \in \mathbb{S}_0^{p,1}(\mathcal{T})$ follow from Lemma 3.7 and the Lax-Milgram Lemma.

As usual, given a basis of the discrete space, the discrete model problem can be rephrased as an equivalent linear system of equations. The bilinear form $a(\cdot, \cdot)$ from

Definition 2.1 and the basis functions $\varphi_n \in \mathbb{S}_0^{p,1}(\mathcal{T})$ from Definition 2.6 compose the governing system matrix.

Definition 2.9 We define the system matrix

$$\boldsymbol{A} := (a(\varphi_n, \varphi_m))_{m,n=1}^{N} \in \mathbb{R}^{N\times N}.$$

Note that the unique solvability of the discrete model problem already ensures that the matrix $\boldsymbol{A}$ is invertible.

2.5 Hierarchical matrices

First, let us sketch the concepts of *cluster trees* and *block cluster trees*. For a more detailed introduction, see, e.g., [21, Chapter 5].

A *(binary) clustering strategy* consists of two mappings child_1, $\mathrm{child}_2 : \mathrm{Pow}\{1,\dots,N\} \longrightarrow \mathrm{Pow}\{1,\dots,N\}$ that satisfy the disjointness property $\mathrm{child}_1(I) \cap \mathrm{child}_2(I) = \emptyset$ and the covering property $\mathrm{child}_1(I) \cup \mathrm{child}_2(I) = I$. (See, e.g., [21] for some examples of such clustering strategies.) Let $I_{\mathrm{root}} := \{1,\dots,N\}$ and $\sigma_{\mathrm{small}} > 0$. Consider a system $\mathbb{T}_N \subseteq \mathrm{Pow}\{1,\dots,N\}$ with $I_{\mathrm{root}} \in \mathbb{T}_N$ that is closed in the following sense: For every tree node $I \in \mathbb{T}_N$ with $\#I > \sigma_{\mathrm{small}}$, its children satisfy $\mathrm{child}_1(I), \mathrm{child}_2(I) \in \mathbb{T}_N$ as well. If $\mathbb{T}_N$ is minimal, i.e., removing one of its elements would violate said properties, then we call $\mathbb{T}_N$ a *(hierarchical) cluster tree*. The assumed minimality allows us to assign a *level* to each tree node such that $\mathrm{level}(I_{\mathrm{root}}) = 0$ and $\mathrm{level}(\mathrm{child}_1(I)) = \mathrm{level}(\mathrm{child}_2(I)) = \mathrm{level}(I) + 1$ for all $I \in \mathbb{T}_N$. Finally, we set $\mathrm{depth}(\mathbb{T}_N) := \max_{I\in\mathbb{T}_N} \mathrm{level}(I)$.

Said clustering strategy induces functions

$$\mathrm{child}_{11}, \mathrm{child}_{12}, \mathrm{child}_{21}, \mathrm{child}_{22} : (\mathrm{Pow}\{1,\dots,N\})^2 \longrightarrow (\mathrm{Pow}\{1,\dots,N\})^2$$

via $\mathrm{child}_{\alpha\beta}(I,J) := (\mathrm{child}_\alpha(I), \mathrm{child}_\beta(J))$. Let $\sigma_{\mathrm{adm}} > 0$. Consider a system $\mathbb{T}_{N\times N} \subseteq (\mathrm{Pow}\{1,\dots,N\})^2$ with $(I_{\mathrm{root}}, I_{\mathrm{root}}) \in \mathbb{T}_{N\times N}$ that is closed in the following sense: For every tree node $(I,J) \in \mathbb{T}_{N\times N}$ with $\mathrm{diam}_{\mathcal{T}}(\mathcal{T}(I)) > \sigma_{\mathrm{adm}} \mathrm{dist}_{\mathcal{T}}(\mathcal{T}(I), \mathcal{T}(J))$, its children satisfy $\mathrm{child}_{11}(I,J)$, $\mathrm{child}_{12}(I,J)$, $\mathrm{child}_{21}(I,J)$, $\mathrm{child}_{22}(I,J) \in \mathbb{T}_{N\times N}$ as well. Again, if $\mathbb{T}_{N\times N}$ is minimal, it is called *(hierarchical) block cluster tree*. The associated *sparsity constant* is given by

$$C_{\mathrm{sparse}}(\mathbb{T}_{N\times N}) := \max\left\{ \max_{I\in\mathbb{T}_N} \#\{J \in \mathbb{T}_N \mid (I,J) \in \mathbb{T}_{N\times N}\}, \max_{J\in\mathbb{T}_N} \#\{I \in \mathbb{T}_N \mid (I,J) \in \mathbb{T}_{N\times N}\} \right\}.$$

Definition 2.10 The subset $\mathbb{P} \subseteq \mathbb{T}_{N\times N}$ of all block cluster tree leaves is called *hierarchical block partition*. We say that $\mathbb{P}$ is *sparse*, if $\mathrm{depth}(\mathbb{T}_N) \lesssim \ln(N)$ and $C_{\mathrm{sparse}}(\mathbb{T}_{N\times N}) \lesssim 1$.

Remark 2.11 For a mesh $\mathcal{T}$ with locally bounded cardinality and a basis $\{\varphi_1,\dots,\varphi_N\} \subseteq \mathbb{S}_0^{p,1}(\mathcal{T})$ with local dual functions $\{\lambda_1,\dots,\lambda_N\} \subseteq L^2(\Omega)$, there

indeed exists a *sparse* hierarchical block partition: The locality assumption on the dual functions allows us to treat their supports as a group of distinct characteristic points in $\mathbb{R}^d$. Thus, we can apply the results from [16]. There, the authors presented a *geometrically balanced* clustering strategy that ensures the upper bounds $\operatorname{depth}(\mathbb{T}_N) \lesssim \ln(h_{\min,\mathcal{T}}^{-1})$ and $C_{\text{sparse}}(\mathbb{T}_{N\times N}) \lesssim 1$ for the resulting trees $\mathbb{T}_N$ and $\mathbb{T}_{N\times N}$. Using the relation $h_{\min,\mathcal{T}} \gtrsim h_{\mathcal{T}}^{\sigma_{\text{card}}}$ from Definition 2.4 for meshes with locally bounded cardinality, we conclude $\operatorname{depth}(\mathbb{T}_N) \lesssim \ln(N)$, i.e., the hierarchical block partition $\mathbb{P}$ is sparse.

From the construction of the block cluster tree it follows that the elements of the hierarchical block partition can be categorized into two groups. More precisely, we can state the following:

Lemma 2.12 *The hierarchical block partition can be decomposed as* $\mathbb{P} = \mathbb{P}_{\text{adm}} \dot\cup \mathbb{P}_{\text{small}}$ *with*

$$\begin{aligned} \forall (I,J) \in \mathbb{P}_{\text{adm}} : \quad 0 \;&<\; \operatorname{diam}_{\mathcal{T}}(\mathcal{T}(I)) \;\le\; \sigma_{\text{adm}} \operatorname{dist}_{\mathcal{T}}(\mathcal{T}(I), \mathcal{T}(J)), \\ \forall (I,J) \in \mathbb{P}_{\text{small}} : \quad & \min\{\#I, \#J\} \;\le\; \sigma_{\text{small}}. \end{aligned}$$

Definition 2.13 Let $\mathbb{P}$ be a sparse hierarchical block partition and $r \in \mathbb{N}$ a given *block rank bound*. We define the set of *$\mathcal{H}$-matrices* by

$$\mathcal{H}(\mathbb{P}, r) := \{\boldsymbol{B} \in \mathbb{R}^{N\times N} \mid \forall (I,J) \in \mathbb{P}_{\text{adm}} : \exists \boldsymbol{X} \in \mathbb{R}^{I\times r}, \boldsymbol{Y} \in \mathbb{R}^{J\times r} : \boldsymbol{B}|_{I\times J} = \boldsymbol{X}\boldsymbol{Y}^T\}.$$

Remark 2.14 By [21, Lemma 6.13], the memory requirement to store an $\mathcal{H}$-matrix $\boldsymbol{B} \in \mathcal{H}(\mathbb{P}, r)$ can be bounded by the quantity $C_{\text{sparse}}(\mathbb{T}_{N\times N})(\sigma_{\text{small}} + r)\operatorname{depth}(\mathbb{T}_N)N$. Inserting the estimates for the cluster tree depth and the sparsity constant of a *sparse* hierarchical block partition (cf. Definition 2.10), we get an overall bound of $\mathcal{O}(rN \ln N)$ for the memory requirement.

2.6 The main result

The following theorem is the main result of the present work. It states that inverses of FEM matrices with meshes of locally bounded cardinality can be approximated at an exponential rate in the block rank by hierarchical matrices.

Theorem 2.15 *Let $\mathcal{T} \subseteq \operatorname{Pow}(\Omega)$ be a mesh of locally bounded cardinality for some $\sigma_{\text{card}} \ge 1$ in the sense of Definition 2.4 and $\{\varphi_1, \dots, \varphi_N\} \subseteq \mathbb{S}_0^{p,1}(\mathcal{T})$ a basis that allows for a system of local dual functions (see Definition 2.6). Let $a(\cdot,\cdot)$ be the elliptic bilinear form from Definition 2.1 and $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ be the corresponding Galerkin stiffness matrix (Definition 2.9). Finally, let $\mathbb{P}$ be a sparse hierarchical block partition as in Definition 2.10. Then, there exists a constant $\sigma_{\text{exp}} > 0$ such that, for every block rank bound $r \in \mathbb{N}$, there exists an $\mathcal{H}$-matrix $\boldsymbol{B} \in \mathcal{H}(\mathbb{P}, r)$ with*

$$\|\boldsymbol{A}^{-1} - \boldsymbol{B}\|_2 \lesssim N^{\sigma_{\text{card}}} \ln(N) \exp(-\sigma_{\text{exp}} r^{1/(d\sigma_{\text{card}}+1)}).$$

Remark 2.16 Comparing the main result with the previous work [11] shows that the parameter σ_{card} of a mesh with locally bounded cardinality additionally appears. For quasi-uniform meshes, as studied in [11], this parameter is given by $\sigma_{\text{card}} = 1$ and Theorem 2.15 reproduces the main result therein. However, for different meshes such as algebraically graded meshes, this parameter reduces the exponential rate of convergence. With our fully discrete method of proof a dependence on mesh parameters can be expected and seems unavoidable. Nonetheless, a numerical example presented at the end of the paper shows that the rate of convergence indeed depends on σ_{card}, albeit the dependence seems to be weaker than in the theoretical bound of Theorem 2.15.

As shown in Sect. 3.2, *uniform* and *algebraically graded* meshes have locally bounded cardinality. In particular, we immediately get the following corollary.

Corollary 2.17 *Let $\mathcal{T} \subseteq \text{Pow}(\Omega)$ be an algebraically graded mesh with grading exponent $\alpha \geq 1$ (see Definition 3.4). Then, Theorem 2.15 holds verbatim with $\sigma_{\text{card}} = \alpha$.*

3 Proof of main result

3.1 Overview

The techniques employed in the proof of our main result are similar to those developed in [11] for uniform meshes. However, some modifications are necessary to deal with the present case of non-uniform meshes $\mathcal{T}$ and (possibly) global basis functions $\varphi_n \in \mathbb{S}_0^{p,1}(\mathcal{T})$. Additionally, we simplify several parts of the previous proof considerably.

(1) Before we begin the proof, we give a motivation for the assumptions made in Definitions 2.4 and 2.6. In Sect. 3.2, we present two types of meshes with locally bounded cardinality, namely *uniform* and *algebraically graded* meshes. The fact that every uniform mesh has locally bounded cardinality will be used during our proof of Theorem 3.31. The locally bounded cardinality of algebraically graded meshes shows that Theorem 2.15 is applicable for algebraically graded meshes in the sense of Definition 3.4.

Then, in Sect. 3.3, we present a practical choice for the dual functions $\lambda_n \in L^2(\Omega)$ from Definition 2.6 for a common choice of basis functions $\varphi_n \in \mathbb{S}_0^{p,1}(\mathcal{T})$. The results from this section guarantee that Theorem 2.15 can be used for many different types of *finite element* bases, including the classic *hat functions*.

(2) The starting point for our proof is an explicit representation formula for $\boldsymbol{A}^{-1}$. Since $\boldsymbol{A}^{-1}$ represents the act of solving the discretized model problem, it is only natural that the corresponding *discrete solution operator* $S_{\mathcal{T}} : L^2(\Omega) \longrightarrow \mathbb{S}_0^{p,1}(\mathcal{T})$ will be involved. Additionally, this endeavor requires the dual functions $\lambda_n \in L^2(\Omega)$ mentioned earlier. We present the explicit formula for $\boldsymbol{A}^{-1}$ at the end of Sect. 3.4.

(3) In Sect. 3.5 we use this formula to go from the "matrix level" to the "function level": Initially, we reduce the problem of approximating $\boldsymbol{A}^{-1}$ as a whole to the problem of approximating $\boldsymbol{A}^{-1}|_{I\times J}$ for each admissible block $(I,J)\in\mathbb{P}_{\mathrm{adm}}$. (The small blocks $\mathbb{P}_{\mathrm{small}}$ are irrelevant for that matter.) As it turns out, this boils down to the following question:

Given admissible clusters $\mathcal{B},\mathcal{D}\subseteq\mathcal{T}$ and a free parameter $L\in\mathbb{N}$, how can we construct a low-dimensional subspace $V_{\mathcal{B},\mathcal{D},L}\subseteq L^2(\Omega)$ that contains a good approximant of $(S_{\mathcal{T}}f)|_{\mathcal{B}}$ for every $f\in L^2(\Omega)$ with $\mathrm{supp}_{\mathcal{T}}(f)\subseteq\mathcal{D}$? More precisely, we want to achieve the bounds (for some fixed $\kappa\geq 1$)

$$\dim V_{\mathcal{B},\mathcal{D},L}\lesssim L^{\kappa},\qquad \inf_{v\in V_{\mathcal{B},\mathcal{D},L}}\|S_{\mathcal{T}}f-v\|_{L^2(\mathcal{B})}\lesssim 2^{-L}\|f\|_{L^2(\mathcal{D})}.$$

The remaining sections will give an answer to this very question. Since the construction of $V_{\mathcal{B},\mathcal{D},L}$ is fairly technical and by no means straightforward, the proof is split into further parts:

(4) As the notation "$V_{\mathcal{B},\mathcal{D},L}$" already suggests, the notion of *locality* plays a prominent role in almost all parts of the proof. This is why we introduce so called *inflated clusters*, *discrete cut-off functions*, and the *discrete cut-off operator* in Sect. 3.6.

(5) In Sect. 3.7 we investigate an important class of functions for our analysis, the spaces of *locally discrete harmonic functions* $\mathbb{S}_{\mathrm{harm}}(\mathcal{B})\subseteq\mathbb{S}_0^{p,1}(\mathcal{T})$. These subspaces have three important properties: First, for certain $f\in L^2(\Omega)$, they contain the image $S_{\mathcal{T}}f$. Second, they are invariant under the influence of their respective discrete cut-off operators. Third, they allow for the *discrete Caccioppoli inequality*, a key ingredient in deriving the asserted error bounds for $V_{\mathcal{B},\mathcal{D},L}$.

(6) Finally, in Sect. 3.8 we construct the *single-* and *multi-step coarsening operators*. For any given $u\in\mathbb{S}_{\mathrm{harm}}(\mathcal{B}^{\delta})$ on the inflated cluster $\mathcal{B}^{\delta}\supseteq\mathcal{B}$, the single-step coarsening operator $Q_{\mathcal{B}}^{\delta}$ produces a "coarse" approximation $Q_{\mathcal{B}}^{\delta}u\in\mathbb{S}_{\mathrm{harm}}(\mathcal{B})$ with a small approximation error on $\mathcal{B}$. This is by far the most intricate part of the proof and puts all the aforementioned concepts to use. Afterwards, the multi-step coarsening operator $Q_{\mathcal{B}}^{\delta,L}$ is just a combination of $L\in\mathbb{N}$ single-step coarsening operators.

(7) In Sect. 3.9 we merely put all the pieces together and finish the proof of Theorem 2.15.

3.2 Examples of meshes with locally bounded cardinality

In this subsection, we present two representatives of meshes with locally bounded cardinality (cf. Definition 2.4): *Uniform* meshes and *algebraically graded* meshes. To verify the locally bounded cardinality property for a given mesh, the following lemma is helpful.

Lemma 3.1 *Let $\mathcal{T} \subseteq \mathrm{Pow}(\Omega)$ be a shape-regular mesh as in Definition 2.2. Then, there hold the bounds*

$$\frac{1}{h_{\mathcal{T}}^d} \lesssim \#\mathcal{T}, \qquad \forall \mathcal{B} \subseteq \mathcal{T}: \quad \#\mathcal{B} \lesssim \left(1 + \frac{\mathrm{diam}_{\mathcal{T}}(\mathcal{B})}{h_{\min,\mathcal{B}}}\right)^d.$$

Proof The first relation follows immediately from $1 \eqsim |\Omega| = \sum_{T\in\mathcal{T}} |T| \eqsim \sum_{T\in\mathcal{T}} h_T^d \leq h_{\mathcal{T}}^d \#\mathcal{T}$. For the second bound, let $\mathcal{B} \subseteq \mathcal{T}$. Using Lemma 3.15 ahead, we get $h_{\min,\mathcal{B}}^d \#\mathcal{B} \leq \sum_{T\in\mathcal{B}} h_T^d \eqsim \sum_{T\in\mathcal{B}} |T| = |\bigcup \mathcal{B}| \lesssim (\mathrm{diam}_{\mathcal{T}}(\mathcal{B}) + h_{\mathcal{B}})^d \lesssim (\mathrm{diam}_{\mathcal{T}}(\mathcal{B}) + h_{\min,\mathcal{B}})^d$. This concludes the proof. □

Definition 3.2 *(Uniform mesh)* A mesh $\mathcal{T} \subseteq \mathrm{Pow}(\Omega)$ is called *uniform*, if there exists a constant $\sigma_{\mathrm{unif}} \geq 1$ such that

$$h_{\min,\mathcal{T}} \leq h_{\mathcal{T}} \leq \sigma_{\mathrm{unif}} h_{\min,\mathcal{T}}.$$

Using Lemma 3.1, we immediately get the following result:

Lemma 3.3 *Every uniform mesh $\mathcal{T} \subseteq \mathrm{Pow}(\Omega)$ has locally bounded cardinality with $\sigma_{\mathrm{card}} = 1$.*

Definition 3.4 (*Mesh graded towards* Γ) Let $\mathcal{T} \subseteq \mathrm{Pow}(\Omega)$ be a mesh and $\Gamma \subseteq \mathbb{R}^d$ satisfy $\Gamma \subseteq \mathbb{R}^d \backslash T$ for all $T \in \mathcal{T}$. Furthermore, let $\alpha \geq 1$ be a *grading exponent* and $H > 0$ a *coarse mesh width*. We say that $\mathcal{T}$ is *(algebraically) graded towards* Γ *with parameters* α, H, if there holds

$$\forall T \in \mathcal{T}: \qquad h_T \eqsim \mathrm{dist}_2(x_T, \Gamma)^{1-1/\alpha} H.$$

Here, x_T denotes the incenter of the element T and $\mathrm{dist}_2(x_T, \Gamma) = \inf_{\gamma\in\Gamma} \|x_T - \gamma\|_2$ is the Euclidean distance between a point and a set.

The set Γ, towards which the mesh is graded, is usually determined by the given problem. For example, reentrant corners of the domain Ω or regions of non-smoothness of the data may entail a reduced regularity of the solution u to the model problem from Sect. 2.1. This usually leads to reduced order of convergence of the finite element approximation on quasiuniform meshes. Choosing the set Γ to contain all singularities of the solution as well as choosing the parameter α correctly, one can restore the optimal order of convergence. To a large extent, the shape of Γ is irrelevant for our analysis. We only require that the mesh resolves Γ, i.e., the mesh can only be graded towards subsets of the mesh skeleton, e.g., vertices/edges/faces in 3D.

Lemma 3.5 *Let $\mathcal{T} \subseteq \mathrm{Pow}(\Omega)$ be a mesh graded towards Γ with parameters α, H. Then, there hold the bounds $H^\alpha \lesssim h_{\min,\mathcal{T}} \leq h_{\mathcal{T}} \lesssim H$. Furthermore, $\mathcal{T}$ has locally bounded cardinality with $\sigma_{\mathrm{card}} = \alpha$.*

Proof We start with the bounds for $h_{\mathcal{T}}$ and $h_{\min,\mathcal{T}}$: For every $T \in \mathcal{T}$, we know from Definition 2.2 that $\mathrm{Ball}_2(x_T, \sigma_{\mathrm{shp}}^{-1} h_T) \subseteq T$. Combining this with the assumption

$\Gamma \subseteq T^c$ from Definition 3.4 yields $\operatorname{dist}_2(x_T, \Gamma) \geq h_T/\sigma_{\mathrm{shp}}$. We conclude $h_T \eqsim \operatorname{dist}_2(x_T, \Gamma)^{1-1/\alpha} H \gtrsim h_T^{1-1/\alpha} H$ and ultimately $h_{\min,\mathcal{T}} \gtrsim H^\alpha$. On the other hand, we have the bound $h_T \eqsim \operatorname{dist}_2(x_T, \Gamma)^{1-1/\alpha} H \leq \sup_{x\in\Omega} \operatorname{dist}_2(x, \Gamma)^{1-1/\alpha} H \lesssim H$ and thus $h_{\mathcal{T}} \lesssim H$.

It remains to prove the locally bounded cardinality: Let $\mathcal{B} \subseteq \mathcal{T}$ be arbitrary. We fix an element $B \in \mathcal{B}$ with $b := \operatorname{dist}_2(x_B, \Gamma) = \min_{T\in\mathcal{B}} \operatorname{dist}_2(x_T, \Gamma)$ and abbreviate $\Delta b := \operatorname{diam}_{\mathcal{T}}(\mathcal{B})$. Note that $h_{\mathcal{B}} \eqsim (\max_{T\in\mathcal{B}} \operatorname{dist}_2(x_T, \Gamma))^{1-1/\alpha} H \lesssim (b + \Delta b)^{1-1/\alpha} H$.

In the case $b \leq \Delta b$ we have the lower bound

$$h_{\min,\mathcal{B}} \geq h_{\min,\mathcal{T}} \gtrsim H^\alpha \gtrsim \frac{h_{\mathcal{B}}^\alpha}{(b + \Delta b)^{\alpha-1}} \geq \frac{h_{\mathcal{B}}^\alpha}{(2\Delta b)^{\alpha-1}}.$$

In the remaining case $b > \Delta b$ we get

$$h_{\min,\mathcal{B}} \eqsim H \left(\min_{T\in\mathcal{B}} \operatorname{dist}_2(x_T, \Gamma) \right)^{1-1/\alpha} = H b^{1-1/\alpha} \gtrsim h_{\mathcal{B}} \left(\frac{b}{b + \Delta b} \right)^{1-1/\alpha} \geq 2^{1/\alpha-1} h_{\mathcal{B}}.$$

In particular, both cases lead to the estimate

$$\#\mathcal{B} \overset{\text{Lem. 3.1}}{\lesssim} \left(1 + \frac{\Delta b}{h_{\min,\mathcal{B}}} \right)^d \lesssim \left(1 + \frac{\Delta b}{h_{\mathcal{B}}} \right)^{d\alpha},$$

which concludes the proof. □

In order to gain a better understanding of Definition 2.4 it is instructive to investigate a counter example as well. To this end, let $\Omega := (0, 1) \subseteq \mathbb{R}$, $M \in \mathbb{N}$, $\xi_m := 2^{-m}$ for all $m \in \{0, \dots, M-1\}$ and $\xi_M := 0$. Then, the elements $T_m := (\xi_m, \xi_{m-1}) \subseteq \Omega$ combine into an *exponentially graded mesh* $\mathcal{T}$ that *does not* have locally bounded cardinality. In fact, $h_{\mathcal{T}} = h_{T_1} = 2^{-1}$ and $h_{\min,\mathcal{T}} = h_{T_M} = 2^{1-M}$, which clearly contradicts the first requirement in Definition 2.4. Finally, note that an exponentially graded mesh can be interpreted as an algebraically graded mesh with grading exponent $\alpha = \infty$ (cf. Definition 3.4).

3.3 Examples of local dual functions

In this subsection, we present a way to construct bases of $\mathbb{S}_0^{p,1}(\mathcal{T})$ that is common in the *finite element* method. This scheme encompasses, in particular, the classical *hat functions* $\varphi_n \in \mathbb{S}_0^{1,1}(\mathcal{T})$ as well as their generalization to $p \geq 1$ (Lagrange elements). Then, we show explicitly how to find a local dual system $\{\lambda_1, \dots, \lambda_N\} \subseteq L^2(\Omega)$ in the sense of Definition 2.6.

Let $p \geq 1$, $L := \dim \mathbb{P}^p(\hat{T})$ and $N := \dim \mathbb{S}_0^{p,1}(\mathcal{T})$. Let $\{\varphi_1, \dots, \varphi_N\} \subseteq \mathbb{S}_0^{p,1}(\mathcal{T})$ be a basis such that:

1) *Local supports:* For every $n \in \{1, \dots, N\}$, there exists an element $T_n \in \mathcal{T}$ such that $T_n \in \operatorname{supp}_{\mathcal{T}}(\varphi_n) \subseteq \mathcal{T}(T_n)$.

2) *Simple structure:* There exists a basis of *shape functions* $\{\hat{\varphi}_1, \dots, \hat{\varphi}_L\} \subseteq \mathbb{P}^p(\hat{T})$ that determines the shape of the basis elements. More precisely, for every

$n \in \{1, \dots, N\}$ and every $T \in \operatorname{supp}_{\mathcal{T}}(\varphi_n)$, there exists an index $\ell(n,T) \in \{1, \dots, L\}$ such that $\varphi_n|_T = \hat{\varphi}_{\ell(n,T)} \circ F_T^{-1}$.

3) *Local distinctness:* The basis functions are *locally distinct* in the following sense: For all $n \neq m \in \{1, \dots, N\}$ and all common $T \in \operatorname{supp}_{\mathcal{T}}(\varphi_n) \cap \operatorname{supp}_{\mathcal{T}}(\varphi_m)$, there holds $\ell(n,T) \neq \ell(m,T)$.

For each basis function φ_n we fix an element $T_n \in \mathcal{T}$ as in 1). Note that a standard scaling argument $T \leftrightarrow \hat{T}$ readily provides the following relation:

$$\forall n \in \{1, \dots, N\} : \qquad \|\varphi_n\|_{L^2(\Omega)} \eqsim h_{T_n}^{d/2}.$$

For the construction of the dual functions $\lambda_n \in L^2(\Omega)$, let $\{\hat{\lambda}_1, \dots, \hat{\lambda}_L\} \subseteq \mathbb{P}^p(\hat{T})$ be the unique set of *dual shape functions*, i.e. $\langle \hat{\varphi}_\ell, \hat{\lambda}_k \rangle_{L^2(\hat{T})} = \delta_{\ell k}$ for all $\ell, k \in \{1, \dots, L\}$. We define the function $\lambda_n \in \mathbb{S}^{p,0}(\mathcal{T}) \subseteq L^2(\Omega)$ in a piecewise manner: For every $T \neq T_n$, we set $\lambda_n|_T := 0$, whereas

$$\lambda_n|_{T_n} := |\det \nabla F_{T_n}|^{-1} \cdot (\hat{\lambda}_{\ell(n,T_n)} \circ F_{T_n}^{-1}).$$

Lemma 3.6 *The subset $\{\lambda_1, \dots, \lambda_N\} \subseteq L^2(\Omega)$ constitutes a local dual system in the sense of Definition 2.6.*

Proof From the definition of λ_n, it is clear that $\operatorname{supp}_{\mathcal{T}}(\lambda_n) = \{T_n\} \subseteq \mathcal{T}(T_n)$. As for the duality, let $n, m \in \{1, \dots, N\}$. If $T_m \notin \operatorname{supp}_{\mathcal{T}}(\varphi_n)$, we have $m \neq n$ and therefore $\langle \varphi_n, \lambda_m \rangle_{L^2(\Omega)} = 0 = \delta_{nm}$. In the remaining case $T_m \in \operatorname{supp}_{\mathcal{T}}(\varphi_n)$ we get

$$\langle \varphi_n, \lambda_m \rangle_{L^2(\Omega)} = \langle \varphi_n, \lambda_m \rangle_{L^2(T_m)} = \langle \hat{\varphi}_{\ell(n,T_m)}, \hat{\lambda}_{\ell(m,T_m)} \rangle_{L^2(\hat{T})} = \delta_{\ell(n,T_m)\ell(m,T_m)} = \delta_{nm}.$$

We compute $|\det \nabla F_{T_m}| = |\hat{T}|^{-1/2} |T_m|^{1/2}$. Recalling that $|T| \eqsim h_T^d$ for every element T in a shape-regular mesh $\mathcal{T}$, we obtain for all $m \in \{1, \dots, N\}$ the relation

$$\begin{aligned} \|\lambda_m\|_{L^2(\Omega)} &= |\det \nabla F_{T_m}|^{-1} \|\hat{\lambda}_{\ell(m,T_m)} \circ F_{T_m}^{-1}\|_{L^2(T_m)} \\ &= |\hat{T}|^{1/2} |T_m|^{-1/2} \|\hat{\lambda}_{\ell(m,T_m)}\|_{L^2(\hat{T})} \eqsim h_{T_m}^{-d/2}. \end{aligned}$$

Finally, for every $T \in \mathcal{T}$, we consider the indices $ms(T) := \{m \mid T_m = T\}$. Due to the duality formula from above, the system $\{\lambda_1, \dots, \lambda_N\} \subseteq \mathbb{S}^{p,0}(\mathcal{T})$ is linearly independent. As a consequence, there must hold $\#ms(T) \lesssim 1$. For every $\boldsymbol{x} \in \mathbb{R}^N$ and every $T \in \mathcal{T}$, we estimate

$$\begin{aligned} \Big\| \sum_{m=1}^N \boldsymbol{x}_m \lambda_m \Big\|_{L^2(T)}^2 &= \Big\| \sum_{m \in ms(T)} \boldsymbol{x}_m \lambda_m \Big\|_{L^2(T)}^2 \\ &\leq \Big(\sum_{m \in ms(T)} \|\lambda_m\|_{L^2(\Omega)}^2 \Big) \Big(\sum_{m \in ms(T)} \boldsymbol{x}_m^2 \Big) \lesssim h_T^{-d} \sum_{m \in ms(T)} \boldsymbol{x}_m^2. \end{aligned}$$

Summing over all elements $T \in \mathcal{T}$ then gives the asserted global stability bound. This concludes the proof. □

3.4 A representation formula for the inverse system matrix

In this subsection, we develop a representation formula for $\boldsymbol{A}^{-1}$ in terms of three linear operators: Recall that $\boldsymbol{A}^{-1}$ represents the action of solving the discrete model problem, so there must be a fundamental connection to the *discrete solution operator* $S_{\mathcal{T}} : L^2(\Omega) \longrightarrow \mathbb{S}_0^{p,1}(\mathcal{T})$. Additionally, we need a way to turn coefficient vectors $\boldsymbol{f} \in \mathbb{R}^N$ into functions $f \in L^2(\Omega)$ that can be plugged into $S_{\mathcal{T}}$. For this purpose, we can use the dual functions $\lambda_n \in L^2(\Omega)$ from Definition 2.6 and the corresponding *coordinate mapping* $\Lambda : \mathbb{R}^N \longrightarrow L^2(\Omega)$. Finally, the image $S_{\mathcal{T}}\Lambda \boldsymbol{f} \in \mathbb{S}_0^{p,1}(\mathcal{T})$ must be converted back to a vector in $\mathbb{R}^N$. A straightforward approach would be to use the inverse $\boldsymbol{\Phi}^{-1}$ of the *coordinate mapping* $\boldsymbol{\Phi} : \mathbb{R}^N \longrightarrow \mathbb{S}_0^{p,1}(\mathcal{T})$ associated with the basis functions $\varphi_n \in \mathbb{S}_0^{p,1}(\mathcal{T})$. But, as it turns out, it is advantageous to use the Hilbert space transpose $\Lambda^T : L^2(\Omega) \longrightarrow \mathbb{R}^N$ instead.

First, let us recall the following classical result:

Lemma 3.7 *The bilinear form $a(\cdot,\cdot)$ from Definition 2.1 is coercive and continuous*:

$$\forall u, v \in H_0^1(\Omega) : \qquad \|u\|_{H^1(\Omega)}^2 \lesssim a(u,u), \qquad |a(u,v)| \lesssim \|u\|_{H^1(\Omega)} \|v\|_{H^1(\Omega)}.$$

The precise definition of the solution operator $S_{\mathcal{T}}$ is given in the following Definition 3.8 and the coordinate mappings $\boldsymbol{\Phi}$ and Λ are defined in Definition 3.9.

Definition 3.8 Let $a : H_0^1(\Omega) \times H_0^1(\Omega) \longrightarrow \mathbb{R}$ be the bilinear form of Definition 2.1. For every $f \in L^2(\Omega)$, denote by $S_{\mathcal{T}} f \in \mathbb{S}_0^{p,1}(\mathcal{T})$ the unique function satisfying the variational equality

$$\forall v \in \mathbb{S}_0^{p,1}(\mathcal{T}) : \qquad a(S_{\mathcal{T}} f, v) = \langle f, v \rangle_{L^2(\Omega)}.$$

The linear mapping $S_{\mathcal{T}} : L^2(\Omega) \longrightarrow \mathbb{S}_0^{p,1}(\mathcal{T})$ is called *discrete solution operator*.

Recall from Sect. 2.4 that existence and uniqueness of $S_{\mathcal{T}} f$ are provided by the Lax-Milgram Lemma. Additionally, there holds the *a priori* bound $\|S_{\mathcal{T}} f\|_{H^1(\Omega)} \lesssim \|f\|_{L^2(\Omega)}$.

Definition 3.9 Let $\{\varphi_1, \dots, \varphi_N\} \subseteq \mathbb{S}_0^{p,1}(\mathcal{T})$ be a basis and $\{\lambda_1, \dots, \lambda_N\} \subseteq L^2(\Omega)$ be a dual system compliant with Definition 2.6. We denote the corresponding coordinate mappings by

$$\boldsymbol{\Phi} : \begin{cases} \mathbb{R}^N & \longrightarrow \quad \mathbb{S}_0^{p,1}(\mathcal{T}) \\ \boldsymbol{x} & \longmapsto \quad \sum_{n=1}^N \boldsymbol{x}_n \varphi_n \end{cases}, \qquad \Lambda : \begin{cases} \mathbb{R}^N & \longrightarrow \quad L^2(\Omega) \\ \boldsymbol{x} & \longmapsto \quad \sum_{n=1}^N \boldsymbol{x}_n \lambda_n \end{cases}.$$

We summarize the most important properties of $\boldsymbol{\Phi}$ and Λ in the following lemma. As usual, we use the notation $\operatorname{supp}(\boldsymbol{x}) := \{n \in \{1, \dots, N\} \mid \boldsymbol{x}_n \neq 0\}$ for the *support* of a vector $\boldsymbol{x} \in \mathbb{R}^N$. Furthermore, recall from Definition 2.8 the notation $\mathcal{T}(I) \subseteq \mathcal{T}$ for all abstract matrix index sets $I \subseteq \{1, \dots, N\}$.

Lemma 3.10 *The Hilbert space transpose of Λ is given by the operator*

$$\Lambda^T : \begin{cases} L^2(\Omega) & \longrightarrow & \mathbb{R}^N \\ v & \longmapsto & (\langle v, \lambda_n \rangle_{L^2(\Omega)})_{n=1}^N \end{cases}.$$

The restriction of Λ^T to the subspace $\mathbb{S}_0^{p,1}(\mathcal{T}) \subseteq L^2(\Omega)$ coincides with the inverse mapping $\boldsymbol{\Phi}^{-1}$. More precisely, for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^N$ and all $v \in \mathbb{S}_0^{p,1}(\mathcal{T})$, there hold the duality/inversion formulae

$$\langle \boldsymbol{\Phi}\boldsymbol{x}, \Lambda \boldsymbol{y} \rangle_{L^2(\Omega)} = \langle \boldsymbol{x}, \boldsymbol{y} \rangle_2, \qquad \Lambda^T \boldsymbol{\Phi}\boldsymbol{x} = \boldsymbol{x}, \qquad \boldsymbol{\Phi}\Lambda^T v = v.$$

Both Λ and Λ^T preserve locality: For all $\boldsymbol{x} \in \mathbb{R}^N$, $v \in L^2(\Omega)$ and $I \subseteq \{1, \dots, N\}$, we have

$$\operatorname{supp}_{\mathcal{T}}(\Lambda \boldsymbol{x}) \subseteq \mathcal{T}(\operatorname{supp}(\boldsymbol{x})), \qquad \|\Lambda^T v\|_{\ell^2(I)} \leq \|\Lambda\| \, \|v\|_{L^2(\mathcal{T}(I))}.$$

Proof The operator Λ^T is indeed the Hilbert space transpose of Λ: For all $v \in L^2(\Omega)$ and $\boldsymbol{x} \in \mathbb{R}^N$, we compute

$$\langle \Lambda^T v, \boldsymbol{x} \rangle_2 = \sum_{n=1}^N \langle v, \lambda_n \rangle_{L^2(\Omega)} \boldsymbol{x}_n = \left\langle v, \sum_{n=1}^N \boldsymbol{x}_n \lambda_n \right\rangle_{L^2(\Omega)} = \langle v, \Lambda \boldsymbol{x} \rangle_{L^2(\Omega)}.$$

The duality formula is a direct consequence of the duality property $\langle \varphi_n, \lambda_m \rangle_{L^2(\Omega)} = \delta_{nm}$ from Definition 2.6: For all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^N$, we have

$$\langle \boldsymbol{\Phi}\boldsymbol{x}, \Lambda \boldsymbol{y} \rangle_{L^2(\Omega)} = \sum_{n,m=1}^N \boldsymbol{x}_n \boldsymbol{y}_m \langle \varphi_n, \lambda_m \rangle_{L^2(\Omega)} = \sum_{n=1}^N \boldsymbol{x}_n \boldsymbol{y}_n = \langle \boldsymbol{x}, \boldsymbol{y} \rangle_2.$$

From this, we immediately get the inversion formula $\Lambda^T \boldsymbol{\Phi}\boldsymbol{x} = \boldsymbol{x}$ as well. On the other hand, for every $v \in \mathbb{S}_0^{p,1}(\mathcal{T})$, there holds $\boldsymbol{\Phi}\Lambda^T v = \boldsymbol{\Phi}\Lambda^T \boldsymbol{\Phi}\boldsymbol{\Phi}^{-1} v = \boldsymbol{\Phi}\boldsymbol{\Phi}^{-1} v = v$.

Next, we turn our attention to the preservation of locality by Λ:

$$\forall \boldsymbol{x} \in \mathbb{R}^N : \qquad \operatorname{supp}_{\mathcal{T}}(\Lambda \boldsymbol{x}) = \operatorname{supp}_{\mathcal{T}}\left(\sum_{n \in \operatorname{supp}(\boldsymbol{x})} \boldsymbol{x}_n \lambda_n \right) \subseteq \bigcup_{n \in \operatorname{supp}(\boldsymbol{x})} \operatorname{supp}_{\mathcal{T}}(\lambda_n)$$
$$\overset{\text{Def. 2.8}}{=} \mathcal{T}(\operatorname{supp}(\boldsymbol{x})).$$

Finally, let $v \in L^2(\Omega)$ and $I \subseteq \{1, \dots, N\}$. Let $\kappa \in L^\infty(\Omega)$ be a (discontinuous) cut-off function with $\kappa|_{\mathcal{T}(I)} \equiv 1$ and $\kappa|_{\mathcal{T} \setminus \mathcal{T}(I)} \equiv 0$. Then,

$$\|\Lambda^T v\|_{\ell^2(I)} = \|\Lambda^T(\kappa v)\|_{\ell^2(I)} \leq \|\Lambda^T(\kappa v)\|_2 \leq \|\Lambda^T\| \, \|\kappa v\|_{L^2(\Omega)} = \|\Lambda\| \, \|v\|_{L^2(\mathcal{T}(I))},$$

which finishes the proof. □

Lemma 3.11 *The system matrix $\boldsymbol{A} \in \mathbb{R}^{N \times N}$ from Definition 2.9, the discrete solution operator $S_{\mathcal{T}} : L^2(\Omega) \longrightarrow \mathbb{S}_0^{p,1}(\mathcal{T})$ from Definition 3.8, and the coordinate mapping $\Lambda : \mathbb{R}^N \longrightarrow L^2(\Omega)$ from Definition 3.9 are related via the representation formula*

$$\forall \boldsymbol{f} \in \mathbb{R}^N : \qquad \boldsymbol{A}^{-1} \boldsymbol{f} = \Lambda^T S_{\mathcal{T}} \Lambda \boldsymbol{f}.$$

Proof First, we establish a relationship between $\boldsymbol{A}$ and a by means of the coordinate mapping $\boldsymbol{\Phi}$:

$$\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^N : \qquad \langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{y} \rangle_2 \overset{\text{Def. 2.9}}{=} \sum_{n,m=1}^{N} a(\varphi_n, \varphi_m) \boldsymbol{x}_n \boldsymbol{y}_m \overset{\text{Def. 3.9}}{=} a(\boldsymbol{\Phi}\boldsymbol{x}, \boldsymbol{\Phi}\boldsymbol{y}).$$

Now, using the duality and inversion formulae from Lemma 3.10, we get

$$\begin{aligned} \forall \boldsymbol{f}, \boldsymbol{y} \in \mathbb{R}^N : \quad \langle \boldsymbol{A}\Lambda^T S_{\mathcal{T}} \Lambda \boldsymbol{f}, \boldsymbol{y} \rangle_2 &= a(\boldsymbol{\Phi}\Lambda^T S_{\mathcal{T}} \Lambda \boldsymbol{f}, \boldsymbol{\Phi}\boldsymbol{y}) = a(S_{\mathcal{T}} \Lambda \boldsymbol{f}, \boldsymbol{\Phi}\boldsymbol{y}) \\ &\overset{\text{Def. 3.8}}{=} \langle \Lambda \boldsymbol{f}, \boldsymbol{\Phi}\boldsymbol{y} \rangle_{L^2(\Omega)} = \langle \boldsymbol{f}, \Lambda^T \boldsymbol{\Phi}\boldsymbol{y} \rangle_2 = \langle \boldsymbol{f}, \boldsymbol{y} \rangle_2. \end{aligned}$$

This readily implies the stated representation formula. □

3.5 Reduction from matrix level to function level

In this subsection, we rephrase the original *matrix* approximation problem as a *function* approximation problem. This will remove the abstract matrix indices $I \subseteq \{1, \dots, N\}$ in favor of element clusters $\mathcal{B} \subseteq \mathcal{T}$. The following lemma facilitates a reduction from the full matrix to the individual matrix blocks.

Lemma 3.12 *Let $\mathbb{P}$ be a sparse hierarchical block partition as in Definition 2.10. Then, there holds the estimate*

$$\forall \boldsymbol{B} \in \mathbb{R}^{N \times N} : \qquad \|\boldsymbol{B}\|_2 \lesssim \ln(N) \cdot \max_{(I,J) \in \mathbb{P}} \|\boldsymbol{B}|_{I \times J}\|_2.$$

Proof In [21, Lemma 6.5.8] (see also [8, 19]), the bound

$$\|\boldsymbol{B}\|_2 \leq C_{\text{sparse}}(\mathbb{T}_{N \times N}) \text{depth}(\mathbb{T}_N) \max_{(I,J) \in \mathbb{P}} \|\boldsymbol{B}|_{I \times J}\|_2$$

was established. Inserting the bounds for the cluster tree depth and sparsity constant from Definition 2.10 immediately gives the desired result. □

The following lemma is the main step in shifting the original problem from matrices to function spaces. Note that the representation formula for $\boldsymbol{A}^{-1}$ from Lemma 3.11 plays a crucial role in its proof.

Lemma 3.13 *Let $(I, J) \in \mathbb{P}_{\text{adm}}$ and $V \subseteq L^2(\Omega)$ be a finite-dimensional subspace. Then, there exist matrices $\boldsymbol{X} \in \mathbb{R}^{I \times r}$ and $\boldsymbol{Y} \in \mathbb{R}^{J \times r}$ with $r \leq \dim V$ such that there holds the error bound*

$$\|\boldsymbol{A}^{-1}|_{I \times J} - \boldsymbol{X}\boldsymbol{Y}^T\|_2 \leq \|\Lambda\|^2 \cdot \sup_{\substack{f \in L^2(\Omega): \\ \text{supp}_{\mathcal{T}}(f) \subseteq \mathcal{T}(J)}} \inf_{v \in V} \frac{\|S_{\mathcal{T}} f - v\|_{L^2(\mathcal{T}(I))}}{\|f\|_{L^2(\Omega)}}.$$

Proof We use the transposed coordinate mapping $\Lambda^T : L^2(\Omega) \longrightarrow \mathbb{R}^N$ from Lemma 3.10 to define $\boldsymbol{V} := (\Lambda^T V)|_I \subseteq \mathbb{R}^I$. Note that $r := \dim \boldsymbol{V} \leq \dim V$. Next, let

the columns of the matrix $\boldsymbol{X} \in \mathbb{R}^{I\times r}$ be an $\ell^2(I)$-orthonormal basis of $\boldsymbol{V}$. In particular, the product $\boldsymbol{X}\boldsymbol{X}^T \in \mathbb{R}^{I\times I}$ represents the $\ell^2(I)$-orthogonal projection from $\mathbb{R}^I$ onto $\boldsymbol{V}$. Finally, set $\boldsymbol{Y} := (\boldsymbol{A}^{-1}|_{I\times J})^T \boldsymbol{X} \in \mathbb{R}^{J\times r}$.

For every $\boldsymbol{f} \in \mathbb{R}^N$ with $\operatorname{supp}(\boldsymbol{f}) \subseteq J$, we get the bound

$$\|(\boldsymbol{A}^{-1}|_{I\times J} - \boldsymbol{X}\boldsymbol{Y}^T)\boldsymbol{f}|_J\|_{\ell^2(I)} = \|(\boldsymbol{I} - \boldsymbol{X}\boldsymbol{X}^T)(\boldsymbol{A}^{-1}\boldsymbol{f})|_I\|_{\ell^2(I)} = \inf_{\boldsymbol{v}\in\boldsymbol{V}} \|(\boldsymbol{A}^{-1}\boldsymbol{f})|_I - \boldsymbol{v}\|_{\ell^2(I)}$$
$$\overset{\text{Lem. 3.11}}{=} \inf_{v\in V} \|\Lambda^T(S_{\mathcal{T}}\Lambda\boldsymbol{f} - v)\|_{\ell^2(I)} \overset{\text{Lem. 3.10}}{\leq} \|\Lambda\| \cdot \inf_{v\in V} \|S_{\mathcal{T}}\Lambda\boldsymbol{f} - v\|_{L^2(\mathcal{T}(I))}.$$

We can divide both sides by $\|\boldsymbol{f}\|_{\ell^2(J)}$, take suprema and substitute $f := \Lambda\boldsymbol{f} \in L^2(\Omega)$. Finally, we use $\operatorname{supp}_{\mathcal{T}}(f) = \operatorname{supp}_{\mathcal{T}}(\Lambda\boldsymbol{f}) \subseteq \mathcal{T}(\operatorname{supp}(\boldsymbol{f})) \subseteq \mathcal{T}(J)$ and $\|\boldsymbol{f}\|_{\ell^2(J)}^{-1} \leq \|\Lambda\|\,\|f\|_{L^2(\Omega)}^{-1}$ to get the desired result. □

A thorough understanding of the preceding lemma is absolutely fundamental for the subsequent sections. Therefore, let us recall its interpretation from Sect. 3.1:

Given $\mathcal{B}, \mathcal{D} \subseteq \mathcal{T}$ with $0 < \operatorname{diam}_{\mathcal{T}}(\mathcal{B}) \leq \sigma_{\text{adm}} \operatorname{dist}_{\mathcal{T}}(\mathcal{B}, \mathcal{D})$ and $L \in \mathbb{N}$, how can we construct a subspace $V_{\mathcal{B},\mathcal{D},L} \subseteq L^2(\Omega)$ of dimension $\dim V_{\mathcal{B},\mathcal{D},L} \lesssim L^\kappa$ (for some fixed $\kappa \geq 1$) that satisfies the error bound

$$\inf_{v\in V_{\mathcal{B},\mathcal{D},L}} \|S_{\mathcal{T}}f - v\|_{L^2(\mathcal{B})} \lesssim 2^{-L}\|f\|_{L^2(\mathcal{D})},$$

for all source functions $f \in L^2(\Omega)$ with $\operatorname{supp}_{\mathcal{T}}(f) \subseteq \mathcal{D}$?

3.6 The discrete cut-off operator

The notion of *cluster inflation* provides a means of enlarging a given cluster by a predefined threshold with respect to the mesh metric $\operatorname{dist}_{\mathcal{T}}(\cdot,\cdot)$ from Definition 2.3. This is one of the core concepts in our proof and will be used extensively. We acknowledge this fact with tight notation:

Definition 3.14 For every cluster $\mathcal{B} \subseteq \mathcal{T}$ and every radius $\delta \geq 0$, we introduce the *inflated cluster*

$$\mathcal{B}^\delta := \{T \in \mathcal{T} \mid \operatorname{dist}_{\mathcal{T}}(T, \mathcal{B}) \leq \delta\}.$$

We summarize the most important facts about the mesh metric and inflated clusters in the subsequent lemma.

Lemma 3.15 *The mesh metric* $\operatorname{dist}_{\mathcal{T}}(\cdot,\cdot)$ *from Definition 2.3 defines a metric on* $\mathcal{T}$. *There holds the triangle type inequality*

$$\forall \mathcal{A}, \mathcal{B}, \mathcal{C} \subseteq \mathcal{T}: \qquad \operatorname{dist}_{\mathcal{T}}(\mathcal{A}, \mathcal{C}) \leq \operatorname{dist}_{\mathcal{T}}(\mathcal{A}, \mathcal{B}) + \operatorname{diam}_{\mathcal{T}}(\mathcal{B}) + \operatorname{dist}_{\mathcal{T}}(\mathcal{B}, \mathcal{C}).$$

For every element $T \in \mathcal{T}$ *and every neighbor* $S \in \mathcal{T}(T)$, [*cf.* (2.1)] *the distance is bounded by* $\operatorname{dist}_{\mathcal{T}}(T, S) \leq \sigma_{\text{shp}} h_T$. *On the other hand, for every* $S \in \mathcal{T}\setminus\{T\}$, *we have*

the lower bound $\operatorname{dist}_{\mathcal{T}}(T,S) \geq \sigma_{\text{shp}}^{-1}(h_T + h_S)$. *Additionally, for every cluster* $\mathcal{B} \subseteq \mathcal{T}$, *there holds* $h_{\mathcal{B}} \leq \max\{h_{\min,\mathcal{B}}, \sigma_{\text{shp}}\operatorname{diam}_{\mathcal{T}}(\mathcal{B})\}$.

When dealing with a second mesh $\mathcal{S} \subseteq \operatorname{Pow}(\Omega)$, *cluster diameters are essentially equivalent*:

$$\forall \mathcal{B} \subseteq \mathcal{T}: \qquad \operatorname{diam}_{\mathcal{S}}(\mathcal{S}(\bigcup \mathcal{B})) \leq \operatorname{diam}_{\mathcal{T}}(\mathcal{B}) + 2h_{\mathcal{B}} + 2h_{\mathcal{S}(\bigcup \mathcal{B})}.$$

Finally, consider clusters $\mathcal{B} \subseteq \mathcal{C} \subseteq \mathcal{T}$ *and inflation radii* $\delta,\ \varepsilon \geq 0$. *Then,* $\mathcal{B} \subseteq \mathcal{B}^{\delta} \subseteq (\mathcal{B}^{\delta})^{\varepsilon} \subseteq \mathcal{B}^{\delta+\varepsilon} \subseteq \mathcal{C}^{\delta+\varepsilon}$. *For the cluster patch* $\mathcal{T}(\mathcal{B})$ *we have the inclusion* $\mathcal{T}(\mathcal{B}) \subseteq \mathcal{B}^{\sigma_{\text{shp}}h_{\mathcal{B}}}$. *Finally,* $\operatorname{diam}_{\mathcal{T}}(\mathcal{B}^{\delta}) \leq \operatorname{diam}_{\mathcal{T}}(\mathcal{B}) + 2\delta$ *and* $h_{\mathcal{B}^{\delta}} \leq \max\{h_{\mathcal{B}}, \sigma_{\text{shp}}\delta\}$.

Proof Both the verification of the metric axioms for $\operatorname{dist}_{\mathcal{T}}(\cdot,\cdot)$ and the triangle type inequality are straightforward.

Let $T \in \mathcal{T}$ and $S \in \mathcal{T}(T)$. From the assumption on shape regularity following Definition 2.2, we know that $x_S \in S \subseteq \bigcup \mathcal{T}(T) \subseteq \operatorname{Ball}_2(x_T, \sigma_{\text{shp}}h_T)$. This yields $\operatorname{dist}_{\mathcal{T}}(T,S) = \|x_S - x_T\|_2 \leq \sigma_{\text{shp}}h_T$. Next, let $T \in \mathcal{T}$ and $S \in \mathcal{T}\setminus\{T\}$. Again, from shape regularity, we know that $\operatorname{Ball}_2(x_S, \sigma_{\text{shp}}^{-1}h_S) \subseteq S$ and $\operatorname{Ball}_2(x_T, \sigma_{\text{shp}}^{-1}h_T) \subseteq T$. Since S and T are disjoint, the inscribed balls have to be disjoint, too. We conclude $\operatorname{dist}_{\mathcal{T}}(T,S) = \|x_S - x_T\|_2 \geq \sigma_{\text{shp}}^{-1}(h_T + h_S)$.

Next, let $\mathcal{B} \subseteq \mathcal{T}$. In the case $\#\mathcal{B} = 1$ we clearly have $h_{\mathcal{B}} = h_{\min,\mathcal{B}}$. In the case $\#\mathcal{B} \geq 2$ we can choose distinct elements $B \neq B_{\max} \in \mathcal{B}$ with $h_{B_{\max}} = h_{\mathcal{B}}$ and conclude $h_{\mathcal{B}} = h_{B_{\max}} \leq \sigma_{\text{shp}}\operatorname{dist}_{\mathcal{T}}(B, B_{\max}) \leq \sigma_{\text{shp}}\operatorname{diam}_{\mathcal{T}}(\mathcal{B})$.

Let $\mathcal{S} \subseteq \operatorname{Pow}(\Omega)$ be an additional mesh, $\mathcal{B} \subseteq \mathcal{T}$ and $B := \bigcup \mathcal{B} \subseteq \mathbb{R}^d$ the associated domain. For every $S \in \mathcal{S}(B)$ there exists a $T(S) \in \mathcal{B}$ with $\overline{T(S)} \cap \overline{S} \neq \emptyset$. Clearly, $\|x_S - x_{T(S)}\|_2 \leq h_S + h_{T(S)}$. We conclude

$$\begin{aligned}\operatorname{diam}_{\mathcal{S}}(\mathcal{S}(B)) &\leq \max_{R,S \in \mathcal{S}(B)} \|x_S - x_{T(S)}\|_2 + \|x_{T(S)} - x_{T(R)}\|_2 + \|x_{T(R)} - x_R\|_2 \\ &\leq \operatorname{diam}_{\mathcal{T}}(\mathcal{B}) + 2h_{\mathcal{B}} + 2h_{\mathcal{S}(B)}.\end{aligned}$$

The inclusion chain for inflated clusters follows directly from Definition 3.14 and the triangle type inequality above.

To see the inclusion of a cluster patch in some inflated cluster, let $\mathcal{B} \subseteq \mathcal{T}$. Then, for every $T \in \mathcal{T}(\mathcal{B})$ there exists a $B \in \mathcal{B}$ with $T \in \mathcal{T}(B)$. We get $\operatorname{dist}_{\mathcal{T}}(T,\mathcal{B}) \leq \operatorname{dist}_{\mathcal{T}}(T,B) \leq \sigma_{\text{shp}}h_B \leq \sigma_{\text{shp}}h_{\mathcal{B}}$, i.e., the inclusion $\mathcal{T}(\mathcal{B}) \subseteq \mathcal{B}^{\sigma_{\text{shp}}h_{\mathcal{B}}}$.

Once again, let $\mathcal{B} \subseteq \mathcal{T}$ and $\delta \geq 0$. The inequality $\operatorname{diam}_{\mathcal{T}}(\mathcal{B}^{\delta}) \leq \operatorname{diam}_{\mathcal{T}}(\mathcal{B}) + 2\delta$ can be derived from the mesh metric's triangle inequality. Finally, to find an upper bound for $h_{\mathcal{B}^{\delta}}$, consider an arbitrary $T \in \mathcal{B}^{\delta}$. By definition, there exists a $B \in \mathcal{B}$ with $\operatorname{dist}_{\mathcal{T}}(T,B) \leq \delta$. In the case $T = B$ we get $h_T = h_B \leq h_{\mathcal{B}}$ and in the remaining case $T \neq B$ we have $h_T \leq \sigma_{\text{shp}}\operatorname{dist}_{\mathcal{T}}(T,B) \leq \sigma_{\text{shp}}\delta$. □

For the construction of the cut-off function $\kappa_{\mathcal{B}}^{\delta}$ in Lemma 3.18 we will use a variant of the classical *Clément operator*, [10].

Definition 3.16 Let $\mathcal{N} \subseteq \overline{\Omega}$ be the nodes of the mesh $\mathcal{T}$ and denote by $\{b_N \,|\, N \in \mathcal{N}\} \subseteq \mathbb{S}^{1,1}(\mathcal{T})$ the well-known *hat-functions*, i.e., $b_N(M) = \delta_{NM}$. We write $\langle v \rangle_T := |T|^{-1} \int_T v \,\mathrm{d}x \in \mathbb{R}$ for the mean value of a function $v \in L^2(\Omega)$ on an element $T \in \mathcal{T}$. Now, the *Clément operator* $J_\mathcal{T} : L^2(\Omega) \longrightarrow \mathbb{S}^{1,1}(\mathcal{T})$ is defined in a nodewise fashion: For every $v \in L^2(\Omega)$, we set $J_\mathcal{T} v := \sum_{N \in \mathcal{N}} \beta_N b_N$, where the nodal value β_N is given by

$$\beta_N := \frac{1}{\#\mathcal{T}(N)} \sum_{T \in \mathcal{T}(N)} \langle v \rangle_T.$$

Lemma 3.17 *The linear operator $J_\mathcal{T}$ has a local projection property*: *Given a cluster $\mathcal{B} \subseteq \mathcal{T}$ and a function $v \in L^2(\Omega)$ with $v|_{\mathcal{T}(\mathcal{B})} \equiv \text{const}$, there holds $(J_\mathcal{T} v)|_\mathcal{B} = v|_\mathcal{B}$. Furthermore, $J_\mathcal{T}$ preserves discrete supports: For every $q \geq 0$ and every $v \in \mathbb{S}^{q,0}(\mathcal{T})$, there holds $\operatorname{supp}_\mathcal{T}(J_\mathcal{T} v) \subseteq \mathcal{T}(\operatorname{supp}(v))$. Moreover, $J_\mathcal{T}$ preserves ranges: For every $v \in \mathbb{S}^{1,0}(\mathcal{T})$ with $0 \leq v \leq 1$ there also holds $0 \leq J_\mathcal{T} v \leq 1$. Finally, we have the stability bound*

$$\forall v \in L^2(\Omega) : \forall T \in \mathcal{T} : \qquad h_T |J_\mathcal{T} v|_{W^{1,\infty}(T)} \lesssim \max_{S \in \mathcal{T}(T)} |\langle v \rangle_T - \langle v \rangle_S|.$$

Proof We only show the stability bound: For every $w \in \mathbb{S}^{1,0}(\mathcal{T})$ and every $T \in \mathcal{T}$, the inverse inequality from Lemma 3.20 provides the estimate $h_T |w|_{W^{1,\infty}(T)} \lesssim \|w\|_{L^\infty(T)} = \max_{N \in \mathcal{N}(T)} |w(N)|$, where $\mathcal{N}(T)$ denotes the nodes of the element T. Since $|\cdot|_{W^{1,\infty}(T)}$ annihilates constants, we also get $h_T |w|_{W^{1,\infty}(T)} \lesssim \max_{N,M \in \mathcal{N}(T)} |w(N) - w(M)|$. Inserting $w := J_\mathcal{T} v \in \mathbb{S}^{1,1}(\mathcal{T})$ and using $(J_\mathcal{T} v)(N) = \beta_N$, the asserted stability bound follows readily. □

The discretized model problem $a(u, v) = \langle f, v \rangle_{L^2(\Omega)}$ was phrased in terms of *global* functions $u, v \in \mathbb{S}^{p,1}_0(\mathcal{T})$. But if we plug in a function v with local support, e.g., $\operatorname{supp}_\mathcal{T}(v) \subseteq \mathcal{B}$ for some prescribed cluster $\mathcal{B} \subseteq \mathcal{T}$, we can extract local information about u on $\mathcal{B}$. This motivates the usage of *discrete cut-off functions*.

Lemma 3.18 *Let $\mathcal{B} \subseteq \mathcal{T}$ and $\delta > 0$ with $4\sigma_{\text{shp}}^3 h_\mathcal{B} \leq \delta \lesssim 1$. Then, there exists a discrete cut-off function $\kappa_\mathcal{B}^\delta$ with*

$$\kappa_\mathcal{B}^\delta \in \mathbb{S}^{1,1}(\mathcal{T}), \quad \operatorname{supp}_\mathcal{T}(\kappa_\mathcal{B}^\delta) \subseteq \mathcal{B}^\delta, \quad \kappa_\mathcal{B}^\delta|_\mathcal{B} \equiv 1, \quad 0 \leq \kappa_\mathcal{B}^\delta \leq 1, \quad \|\kappa_\mathcal{B}^\delta\|_{W^{1,\infty}(\Omega)} \lesssim \frac{1}{\delta}.$$

Proof We abbreviate $\varepsilon := \delta/(4\sigma_{\text{shp}}^2) > 0$ and consider a step function $\kappa \in \mathbb{S}^{0,0}(\mathcal{T})$ defined by

$$\forall T \in \mathcal{T} : \qquad \kappa|_T := \max\{0, 1 - \operatorname{dist}_\mathcal{T}(T, \mathcal{T}(\mathcal{B}))/\varepsilon\} \in \mathbb{R}.$$

From the definition we immediately get $\operatorname{supp}_\mathcal{T}(\kappa) \subseteq \mathcal{T}(\mathcal{B})^\varepsilon$ and $\kappa|_{\mathcal{T}(\mathcal{B})} \equiv 1$ as well as $0 \leq \kappa \leq 1$. (Recall that $\mathcal{T}(\mathcal{B})$ are all patch elements of $\mathcal{B}$ and $\mathcal{T}(\mathcal{B})^\varepsilon$ is the corresponding inflated cluster by a radius of ε.) Next, for every $T \in \mathcal{T}$ and every neighbor $S \in \mathcal{T}(T)$, we apply the triangle inequality from Lemma 3.15 to the clusters

$\{T\}, \{S\}, \mathcal{T}(\mathcal{B})$ and derive $\operatorname{dist}_{\mathcal{T}}(T, \mathcal{T}(\mathcal{B})) \le \operatorname{dist}_{\mathcal{T}}(T,S) + \operatorname{dist}_{\mathcal{T}}(S, \mathcal{T}(\mathcal{B}))$. (Recall from Definition 2.3 that $\operatorname{diam}_{\mathcal{T}}(S) = 0$, since $\{S\}$ contains only one element.) Exploiting the Lipschitz continuity of $t \mapsto \max\{0,t\}$, we get the error bound

$$|\kappa|_T - \kappa|_S| \le \frac{|\operatorname{dist}_{\mathcal{T}}(T, \mathcal{T}(\mathcal{B})) - \operatorname{dist}_{\mathcal{T}}(S, \mathcal{T}(\mathcal{B}))|}{\varepsilon} \le \frac{\operatorname{dist}_{\mathcal{T}}(T,S)}{\varepsilon} \overset{\text{Lem. 3.15}}{\lesssim} \frac{h_T}{\varepsilon} \eqsim \frac{h_T}{\delta}.$$

We use the Clément operator $J_{\mathcal{T}} : L^2(\Omega) \longrightarrow \mathbb{S}^{1,1}(\mathcal{T})$ from Definition 3.16 to define $\kappa_{\mathcal{B}}^{\delta} := J_{\mathcal{T}}\kappa \in \mathbb{S}^{1,1}(\mathcal{T})$. For the support of $\kappa_{\mathcal{B}}^{\delta}$ we compute

$$\begin{aligned}\operatorname{supp}_{\mathcal{T}}(\kappa_{\mathcal{B}}^{\delta}) &\overset{\text{Lem. 3.17}}{\subseteq} \mathcal{T}(\operatorname{supp}_{\mathcal{T}}(\kappa)) \subseteq \mathcal{T}(\mathcal{T}(\mathcal{B})^{\varepsilon}) \overset{\text{Lem. 3.15}}{\subseteq} \mathcal{B}^{(1+\sigma_{\mathrm{shp}}^2)(\sigma_{\mathrm{shp}} h_{\mathcal{B}} + \varepsilon)} \\ &\subseteq \mathcal{B}^{2\sigma_{\mathrm{shp}}^3 h_{\mathcal{B}} + \delta/2} \subseteq \mathcal{B}^{\delta},\end{aligned}$$

where in the last step we used $\delta \ge 4\sigma_{\mathrm{shp}}^3 h_{\mathcal{B}}$.

From Lemma 3.17 and $\kappa|_{\mathcal{T}(\mathcal{B})} \equiv 1$ we get $\kappa_{\mathcal{B}}^{\delta}|_{\mathcal{B}} \equiv 1$. Moreover, $0 \le \kappa \le 1$ yields $0 \le \kappa_{\mathcal{B}}^{\delta} \le 1$. This implies, in particular, $\|\kappa_{\mathcal{B}}^{\delta}\|_{L^\infty(\Omega)} \le 1 \lesssim \delta^{-1}$, where we used the assumption $\delta \lesssim 1$. The remaining bound $|\kappa_{\mathcal{B}}^{\delta}|_{W^{1,\infty}(\Omega)} \lesssim \delta^{-1}$ follows from

$$\forall T \in \mathcal{T}: \qquad h_T |\kappa_{\mathcal{B}}^{\delta}|_{W^{1,\infty}(T)} \overset{\text{Lem. 3.17}}{\lesssim} \max_{S \in \mathcal{T}(T)} |\kappa|_T - \kappa|_S| \lesssim \frac{h_T}{\delta}.$$

This finishes the proof. □

Given a cluster $\mathcal{B} \subseteq \mathcal{T}$ and a distance $\delta > 0$, the discrete cut-off function $\kappa_{\mathcal{B}}^{\delta}$ allows us to "restrict" a function $v \in \mathbb{S}^{p,1}(\mathcal{T})$ to the subdomain $\bigcup \mathcal{B}^{\delta} \subseteq \Omega$ while preserving continuity. This can be achieved by simply multiplying v with $\kappa_{\mathcal{B}}^{\delta}$. Note that the product $\kappa_{\mathcal{B}}^{\delta} v$ has polynomial degree $p+1$, rather than p. To mitigate this drawback, we can simply re-interpolate the result with an operator of order p.

Definition 3.19 Let $p \ge 1$ and denote by $\hat{I}^p : C^0(\overline{\hat{T}}) \longrightarrow \mathbb{P}^p(\hat{T})$ the *(local) Lagrange interpolation operator* on the reference element $\hat{T}$. The *(global) Lagrange interpolation operator* $I_{\mathcal{T}}^p : C_{\mathrm{pw}}^0(\mathcal{T}) \longrightarrow \mathbb{S}^{p,0}(\mathcal{T})$ is defined in a piecewise manner: For every $v \in C_{\mathrm{pw}}^0(\mathcal{T})$ and every $T \in \mathcal{T}$, we set

$$(I_{\mathcal{T}}^p v)|_T := \hat{I}^p (v \circ F_T) \circ F_T^{-1}.$$

In order to derive a useful stability estimate for $I_{\mathcal{T}}^p$, we use a standard element-wise inverse inequality, which follows from scaling arguments.

Lemma 3.20 *Let $k, \ell \in \mathbb{N}_0$ with $k \ge \ell \ge 0$, $q \in [1,\infty]$ and $p \ge 0$. Then, for all discrete functions $v \in \mathbb{S}^{p,0}(\mathcal{T})$ and all elements $T \in \mathcal{T}$, there holds the inverse inequality*

$$h_T^k |v|_{W^{k,q}(T)} \lesssim h_T^{\ell} |v|_{W^{\ell,q}(T)}.$$

The properties of the Lagrange interpolation operator $I_{\mathcal{T}}^p$ are very similar to those of the Clément operator $J_{\mathcal{T}}$ from Definition 3.16. For the sake of completeness, we include them in the following lemma.

Lemma 3.21 *Let $p \geq 1$. The linear operator $I_{\mathcal{T}}^p$ has a local projection property: Given a cluster $\mathcal{B} \subseteq \mathcal{T}$ and a function $v \in C_{\mathrm{pw}}^0(\mathcal{T})$ with $v \in \mathbb{S}^{p,0}(\mathcal{B})$, there holds $(I_{\mathcal{T}}^p v)|_{\mathcal{B}} = v|_{\mathcal{B}}$. Furthermore, $I_{\mathcal{T}}^p$ preserves global continuity and homogeneous boundary values: For every $v \in C^0(\overline{\Omega})$, there holds $I_{\mathcal{T}}^p v \in \mathbb{S}^{p,1}(\mathcal{T})$. Similarly, if $v \in C^0(\overline{\Omega})$ with $v|_{\partial\Omega} \equiv 0$, then $I_{\mathcal{T}}^p v \in \mathbb{S}_0^{p,1}(\mathcal{T})$. Moreover, $I_{\mathcal{T}}^p$ preserves discrete supports: For every $q \geq 0$ and every $v \in \mathbb{S}^{q,0}(\mathcal{T})$, we have $\operatorname{supp}_{\mathcal{T}}(I_{\mathcal{T}}^p v) \subseteq \operatorname{supp}_{\mathcal{T}}(v)$. Finally, for all $q \geq 0$, $v \in \mathbb{S}^{q,0}(\mathcal{T})$ and $T \in \mathcal{T}$, there hold the following stability and error estimates (with constants depending on q):*

$$\forall m \in \{0, \dots, p+1\} : \qquad \begin{aligned} |I_{\mathcal{T}}^p v|_{H^m(T)} &\lesssim |v|_{H^m(T)}, \\ \textstyle\sum_{\ell=0}^{p+1} h_T^\ell |(\mathrm{id} - I_{\mathcal{T}}^p)(v)|_{H^\ell(T)} &\lesssim h_T^{p+1} |v|_{H^{p+1}(T)}. \end{aligned}$$

Proof We briefly sketch the proof of the stability and error bounds: The mapping $v \mapsto \|\hat{I}^p v\|_{L^2(\hat{T})} + |v|_{H^{p+1}(\hat{T})}$ defines a norm on the finite-dimensional space $\mathbb{P}^q(\hat{T})$. Hence, by norm equivalence, $\|v\|_{H^{p+1}(\hat{T})} \lesssim \|\hat{I}^p v\|_{L^2(\hat{T})} + |v|_{H^{p+1}(\hat{T})}$ for all $v \in \mathbb{P}^q(\hat{T})$. Inserting $v := w - \hat{I}^p w$ for arbitrary $w \in \mathbb{P}^q(\hat{T})$ results in the bound $\|w - \hat{I}^p w\|_{H^{p+1}(\hat{T})} \lesssim |w|_{H^{p+1}(\hat{T})}$. Finally, a standard scaling argument $\hat{T} \leftrightarrow T$ yields the desired error estimate on T. As for the stability bound, we perform a straightforward triangle inequality on T, reuse the already proven error bound and finish off with the inverse inequality from Lemma 3.20. □

Remark 3.22 The fact that $I_{\mathcal{T}}^p$ preserves global continuity and homogeneous boundary values hinges on an implicit assumption about the (local) interpolation points used by the local Lagrange interpolation operator $\hat{I}^p$. Recall from Definition 2.2 that the reference element $\hat{T} \subseteq \mathbb{R}^d$ is a simplex and thus delimited by $d+1$ hyperplanes. The interpolation points on each hyperplane $\hat{E}$ must be unisolvent for the space $\mathbb{P}^p(\hat{E})$. Then, in particular, every polynomial $v \in \mathbb{P}^p(\hat{T})$ vanishing at the interpolation points in $\hat{E}$ must already vanish everywhere on $\hat{E}$. This property readily implies that homogeneous boundary values are preserved by the global operator $I_{\mathcal{T}}^p$. Finally, the distribution of interpolation points on each hyperplane $\hat{E}$ must be "symmetric". More precisely, if two elements $T_1, T_2 \in \mathcal{T}$ share a common hyperplane, we require the corresponding interpolation points to align perfectly. In this case, using the same argument as before, the operator $I_{\mathcal{T}}^p$ preserves global continuity indeed.

As our next step, we encapsulate the aforementioned "cut-off" process in a linear operator.

Definition 3.23 Let $\mathcal{B} \subseteq \mathcal{T}$ and $\delta > 0$ with $4\sigma_{\mathrm{shp}}^3 h_{\mathcal{B}} \leq \delta \lesssim 1$ and denote by $\kappa_{\mathcal{B}}^\delta \in \mathbb{S}^{1,1}(\mathcal{T})$ the discrete cut-off function from Lemma 3.18. Furthermore, denote by $I_{\mathcal{T}}^p : C_{\mathrm{pw}}^0(\mathcal{T}) \longrightarrow \mathbb{S}^{p,0}(\mathcal{T})$ the Lagrange interpolation operator from Definition 3.19. We define the *discrete cut-off operator*

$$K_{\mathcal{B}}^{\delta} : \begin{cases} \mathbb{S}^{p,1}(\mathcal{T}) & \longrightarrow \mathbb{S}^{p,1}(\mathcal{T}) \\ v & \longmapsto I_{\mathcal{T}}^{p}(\kappa_{\mathcal{B}}^{\delta} v) \end{cases}.$$

The discrete cut-off operator $K_{\mathcal{B}}^{\delta}$ inherits its core properties from $I_{\mathcal{T}}^{p}$.

Lemma 3.24 *Let $\mathcal{B} \subseteq \mathcal{T}$ and $\delta > 0$ with $4\sigma_{\mathrm{shp}}^{3} h_{\mathcal{B}} \leq \delta \lesssim 1$. For all $v \in \mathbb{S}^{p,1}(\mathcal{T})$, the linear operator $K_{\mathcal{B}}^{\delta}$ has the cut-off property $\mathrm{supp}_{\mathcal{T}}(K_{\mathcal{B}}^{\delta} v) \subseteq \mathcal{B}^{\delta}$ and the local projection property $(K_{\mathcal{B}}^{\delta} v)|_{\mathcal{B}} = v|_{\mathcal{B}}$. Furthermore, $K_{\mathcal{B}}^{\delta}$ preserves homogeneous boundary values: For all $v \in \mathbb{S}_0^{p,1}(\mathcal{T})$, there holds $K_{\mathcal{B}}^{\delta} v \in \mathbb{S}_0^{p,1}(\mathcal{T})$. Finally, for every $v \in \mathbb{S}^{p,1}(\mathcal{T})$ and every $T \in \mathcal{T}$, there holds the local stability estimate*

$$\|K_{\mathcal{B}}^{\delta} v\|_{L^2(T)} + \delta |K_{\mathcal{B}}^{\delta} v|_{H^1(T)} \lesssim \|v\|_{L^2(T)} + \delta |v|_{H^1(T)}.$$

Proof The cut-off property, the local projection property and the preservation of homogeneous boundary values follow directly from Lemma 3.21 and Lemma 3.18. Finally, let $v \in \mathbb{S}^{p,1}(\mathcal{T})$ and $T \in \mathcal{T}$. Note that $\kappa_{\mathcal{B}}^{\delta} v \in \mathbb{S}^{p+1,1}(\mathcal{T})$, i.e., we can use the stability estimate from Lemma 3.21:

$$\begin{aligned} \sum_{\ell=0}^{1} \delta^{\ell} |K_{\mathcal{B}}^{\delta} v|_{H^{\ell}(T)} &\lesssim \sum_{\ell=0}^{1} \delta^{\ell} |\kappa_{\mathcal{B}}^{\delta} v|_{H^{\ell}(T)} \lesssim \sum_{\ell=0}^{1} \delta^{\ell} \sum_{i=0}^{\ell} |\kappa_{\mathcal{B}}^{\delta}|_{W^{\ell-i,\infty}(T)} |v|_{H^{i}(T)} \\ &\overset{\text{Lem. 3.18}}{\lesssim} \sum_{\ell=0}^{1} \delta^{\ell} \sum_{i=0}^{\ell} \delta^{i-\ell} |v|_{H^{i}(T)} \lesssim \sum_{\ell=0}^{1} \delta^{\ell} |v|_{H^{\ell}(T)}. \end{aligned}$$

This finishes the proof. □

3.7 The spaces of locally discrete harmonic functions

In this subsection, we introduce the spaces of *locally discrete harmonic functions*. As we already mentioned in Sect. 3.1, they are chosen for three main reasons: To begin with, they fit in seamlessly with the discrete solution operator $S_{\mathcal{T}} : L^2(\Omega) \longrightarrow \mathbb{S}_0^{p,1}(\mathcal{T})$ from Definition 3.8. Furthermore, as specified in Lemma 3.26, they are invariant with respect to the discrete cut-off operators $K_{\mathcal{B}}^{\delta} : \mathbb{S}^{p,1}(\mathcal{T}) \longrightarrow \mathbb{S}^{p,1}(\mathcal{T})$ from Definition 3.23. But most importantly, they contain functions whose H^1-norms can be bounded by L^2-norms with constants independent of h, i.e., a *discrete Caccioppoli inequality* holds.

Definition 3.25 For every $\mathcal{B} \subseteq \mathcal{T}$, the space of *locally discrete harmonic functions* is given by

$$\mathbb{S}_{\mathrm{harm}}(\mathcal{B}) := \{u \in \mathbb{S}_0^{p,1}(\mathcal{T}) \,|\, \forall v \in \mathbb{S}_0^{p,1}(\mathcal{T}) \text{ with } \mathrm{supp}_{\mathcal{T}}(v) \subseteq \mathcal{B} : a(u,v) = 0\} \subseteq \mathbb{S}_0^{p,1}(\mathcal{T}).$$

We summarize the first two main features of the spaces $\mathbb{S}_{\text{harm}}(\mathcal{B})$ in the next lemma, namely, their relationships to the discrete solution operator $S_{\mathcal{T}} : L^2(\Omega) \longrightarrow \mathbb{S}_0^{p,1}(\mathcal{T})$ and the discrete cut-off operators $K_{\mathcal{B}}^{\delta} : \mathbb{S}^{p,1}(\mathcal{T}) \longrightarrow \mathbb{S}^{p,1}(\mathcal{T})$.

Lemma 3.26 *The spaces of locally discrete harmonic functions are nested in the sense*

$$\forall \mathcal{B} \subseteq \mathcal{B}^+ \subseteq \mathcal{T} : \qquad \mathbb{S}_{\text{harm}}(\mathcal{B}^+) \subseteq \mathbb{S}_{\text{harm}}(\mathcal{B}).$$

Furthermore, for all clusters $\mathcal{B}, \mathcal{D} \subseteq \mathcal{T}$ with $\mathcal{B} \cap \mathcal{D} = \emptyset$, the operator $S_{\mathcal{T}}$ has the mapping property

$$\forall f \in L^2(\Omega) \text{ with } \operatorname{supp}_{\mathcal{T}}(f) \subseteq \mathcal{D} : \qquad S_{\mathcal{T}} f \in \mathbb{S}_{\text{harm}}(\mathcal{B}).$$

Finally, for all $\mathcal{B} \subseteq \mathcal{T}$ and all $\delta > 0$ with $4\sigma_{\text{shp}}^3 h_{\mathcal{B}} \leq \delta \lesssim 1$, we have the invariance

$$\forall u \in \mathbb{S}_{\text{harm}}(\mathcal{B}) : \qquad K_{\mathcal{B}}^{\delta} u \in \mathbb{S}_{\text{harm}}(\mathcal{B}).$$

Proof The inclusion $\mathbb{S}_{\text{harm}}(\mathcal{B}^+) \subseteq \mathbb{S}_{\text{harm}}(\mathcal{B})$ follows directly from the definition of the spaces. As for the mapping properties of $S_{\mathcal{T}}$, let $f \in L^2(\Omega)$ with $\operatorname{supp}_{\mathcal{T}}(f) \subseteq \mathcal{D}$. Then, for every $v \in \mathbb{S}_0^{p,1}(\mathcal{T})$ with $\operatorname{supp}_{\mathcal{T}}(v) \subseteq \mathcal{B}$, we have

$$a(S_{\mathcal{T}} f, v) \overset{\text{Def. 3.8}}{=} \langle f, v \rangle_{L^2(\mathcal{D} \cap \mathcal{B})} \overset{\mathcal{B} \cap \mathcal{D} = \emptyset}{=} 0.$$

Finally, consider a function $u \in \mathbb{S}_{\text{harm}}(\mathcal{B})$ and an arbitrary $v \in \mathbb{S}_0^{p,1}(\mathcal{T})$ with $\operatorname{supp}_{\mathcal{T}}(v) \subseteq \mathcal{B}$. Then,

$$\begin{aligned} a(K_{\mathcal{B}}^{\delta} u, v) &\overset{\text{Def. 2.1}}{=} \langle a_1 \nabla K_{\mathcal{B}}^{\delta} u, \nabla v \rangle_{L^2(\mathcal{B})} + \langle a_2 \cdot \nabla K_{\mathcal{B}}^{\delta} u, v \rangle_{L^2(\mathcal{B})} + \langle a_3 K_{\mathcal{B}}^{\delta} u, v \rangle_{L^2(\mathcal{B})} \\ &\overset{\text{Lem. 3.24}}{=} \langle a_1 \nabla u, \nabla v \rangle_{L^2(\mathcal{B})} + \langle a_2 \cdot \nabla u, v \rangle_{L^2(\mathcal{B})} + \langle a_3 u, v \rangle_{L^2(\mathcal{B})} \\ &= a(u, v) \\ &= 0. \end{aligned}$$

This gives $K_{\mathcal{B}}^{\delta} u \in \mathbb{S}_{\text{harm}}(\mathcal{B})$, which concludes the proof. □

Next, we turn our attention to the *discrete Caccioppoli inequality*. In a nutshell, it will allow us to bound an H^1-norm on a cluster $\mathcal{B} \subseteq \mathcal{T}$ by an L^2-norm on the slightly larger cluster $\mathcal{B}^{\delta}$. Obviously, this can only be true for a certain subspace $V \subseteq \mathbb{S}^{p,1}(\mathcal{T})$. In our setting, this is the space of locally discrete harmonic functions $\mathbb{S}_{\text{harm}}(\mathcal{B}^{\delta})$ from Definition 3.25. We can interpret the discrete Caccioppoli inequality as an improved version of the inverse inequality from Lemma 3.20, which bounds an H^1-seminorm by an L^2-norm, too. This time, however, the prefactor h of the H^1-seminorm can be increased to a (possibly much) bigger parameter $\delta \gg h$.

Lemma 3.27 *Let $\mathcal{B} \subseteq \mathcal{T}$ and $\delta > 0$ with $4\sigma_{\text{shp}}^3 h_{\mathcal{B}} \leq \delta \lesssim 1$. Then, for every $u \in \mathbb{S}_{\text{harm}}(\mathcal{B}^{\delta})$, there holds the discrete Caccioppoli inequality*

$$\delta |u|_{H^1(\mathcal{B})} \lesssim \|u\|_{L^2(\mathcal{B}^{\delta})}.$$

Proof First off, by induction on $p \geq 1$ we show the following estimate: For every $\kappa \in \mathbb{S}^{1,1}(\mathcal{T})$, $u \in \mathbb{S}^{p,0}(\mathcal{T})$ and $T \in \mathcal{T}$,

$$h_T^{p+1} |\kappa^2 u|_{H^{p+1}(T)} \lesssim h_T^2 |\kappa|_{W^{1,\infty}(T)} (\|u \nabla \kappa\|_{L^2(T)} + \|\kappa \nabla u\|_{L^2(T)}). \tag{3.1}$$

In the base case $p = 1$, the second order derivatives in $|\kappa^2 u|_{H^2(T)}$ can be computed explicitly. Since κ, $u \in \mathbb{P}^1(T)$, the terms containing $\mathrm{D}^\alpha \kappa$ or $\mathrm{D}^\alpha u$ with $|\alpha| = 2$ are not present. In the induction step $p \mapsto p+1$, we estimate $|\kappa^2 u|_{H^{p+2}(T)} \lesssim \sum_i |\kappa(\partial_i \kappa) u|_{H^{p+1}(T)} + |\kappa^2 (\partial_i u)|_{H^{p+1}(T)}$. For the first summand, we use the inverse inequality Lemma 3.20 and get $|\kappa(\partial_i \kappa) u|_{H^{p+1}(T)} \lesssim h_T^{-p} |\kappa(\partial_i \kappa) u|_{H^1(T)}$. Again, we can expand the derivatives explicitly and cancel all terms containing second order derivatives of $\kappa \in \mathbb{P}^1(T)$. The second summand is directly amenable to the induction hypothesis, i.e., (3.1): $|\kappa^2 (\partial_i u)|_{H^{p+1}(T)} \lesssim h_T^{1-p} |\kappa|_{W^{1,\infty}(T)} (\|(\partial_i u) \nabla \kappa\|_{L^2(T)} + \|\kappa \nabla (\partial_i u)\|_{L^2(T)})$. Since $\nabla \kappa \equiv \text{const}$, we have

$$\|(\partial_i u) \nabla \kappa\|_{L^2(T)} = |\nabla \kappa| \|\partial_i u\|_{L^2(T)} \lesssim h_T^{-1} |\nabla \kappa| \|u\|_{L^2(T)} = h_T^{-1} \|u \nabla \kappa\|_{L^2(T)}.$$

The term $\|\kappa \nabla (\partial_i u)\|_{L^2(T)}$ can be treated with the identity $\kappa \nabla (\partial_i u) = \partial_i (\kappa \nabla u) - (\partial_i \kappa) \nabla u$ and the inverse inequality Lemma 3.20 using the same arguments. Multiplication with h_T^{p+2} then proves the induction step.

Let us turn our attention to the discrete Caccioppoli inequality itself. To this end, let $\mathcal{B} \subseteq \mathcal{T}$ and $\delta > 0$ with $4\sigma_{\mathrm{shp}}^3 h_{\mathcal{B}} \leq \delta \lesssim 1$. We denote by $\kappa := \kappa_{\mathcal{B}}^{\delta} \in \mathbb{S}^{1,1}(\mathcal{T})$ the discrete cut-off function from Lemma 3.18 and by $I_{\mathcal{T}}^p : C_{\mathrm{pw}}^0(\mathcal{T}) \longrightarrow \mathbb{S}^{p,0}(\mathcal{T})$ the Lagrange interpolation operator from Definition 3.19. Furthermore, let $u \in \mathbb{S}_{\mathrm{harm}}(\mathcal{B}^\delta)$. The key step of the proof is to exploit the orthogonality $a(u, v) = 0$ for some carefully chosen test function $v \in \mathbb{S}_0^{p,1}(\mathcal{T})$ with $\mathrm{supp}_{\mathcal{T}}(v) \subseteq \mathcal{B}^\delta$. From Lemmas 3.21 and 3.18 we know that $v := I_{\mathcal{T}}^p(\kappa^2 u)$ satisfies both $v \in \mathbb{S}_0^{p,1}(\mathcal{T})$ and $\mathrm{supp}_{\mathcal{T}}(v) \subseteq \mathrm{supp}_{\mathcal{T}}(\kappa) \subseteq \mathcal{B}^\delta$, i.e., we can use v as a test function. This results in the following bound:

$$\begin{aligned}
a(u, \kappa^2 u) &= a(u, \kappa^2 u - v) = a(u, (\mathrm{id} - I_{\mathcal{T}}^p)(\kappa^2 u)) \\
&\overset{\text{Def. 2.1}}{\lesssim} \sum_{T \in \mathcal{B}^\delta} \|u\|_{H^1(T)} \|(\mathrm{id} - I_{\mathcal{T}}^p)(\kappa^2 u)\|_{H^1(T)} \\
&\overset{\text{Lem. 3.21}}{\lesssim} \sum_{T \in \mathcal{B}^\delta} \|u\|_{H^1(T)} h_T^p |\kappa^2 u|_{H^{p+1}(T)} \\
&\lesssim |\kappa|_{W^{1,\infty}(\Omega)} \sum_{T \in \mathcal{B}^\delta} h_T \|u\|_{H^1(T)} (\|u \nabla \kappa\|_{L^2(T)} + \|\kappa \nabla u\|_{L^2(T)}) \\
&\overset{\text{Lem. 3.20}}{\lesssim} |\kappa|_{W^{1,\infty}(\Omega)} \sum_{T \in \mathcal{B}^\delta} \|u\|_{L^2(T)} (\|u \nabla \kappa\|_{L^2(T)} + \|\kappa \nabla u\|_{L^2(T)}) \\
&\overset{\text{C.Sch.}}{\lesssim} |\kappa|_{W^{1,\infty}(\Omega)} \|u\|_{L^2(\mathcal{B}^\delta)} (\|u \nabla \kappa\|_{L^2(\Omega)} + \|\kappa \nabla u\|_{L^2(\Omega)}).
\end{aligned}$$

On the other hand, using the coercivity of the PDE coefficient a_1 in the bilinear form $a(\cdot,\cdot)$, cf. Sect. 2.1, we can expand the term $a(u,\kappa^2 u)$ and rearrange the summands:

$$\begin{aligned}\|\kappa\nabla u\|_{L^2(\Omega)}^2 &\lesssim \langle a_1\kappa\nabla u,\kappa\nabla u\rangle_{L^2(\Omega)}\\ &\overset{\text{Def. 2.1}}{=} a(u,\kappa^2u) - 2\langle a_1\kappa\nabla u,u\nabla\kappa\rangle_{L^2(\Omega)}\\ &\quad - \langle a_2\cdot\nabla u,\kappa^2u\rangle_{L^2(\Omega)} - \langle a_3u,\kappa^2u\rangle_{L^2(\Omega)}\\ &\lesssim |\kappa|_{W^{1,\infty}(\Omega)}\|u\|_{L^2(\mathcal{B}^\delta)}(\|u\nabla\kappa\|_{L^2(\Omega)} + \|\kappa\nabla u\|_{L^2(\Omega)})\\ &\quad + \|\kappa\nabla u\|_{L^2(\Omega)}\|u\nabla\kappa\|_{L^2(\Omega)} + \|\kappa\nabla u\|_{L^2(\Omega)}\|\kappa u\|_{L^2(\Omega)} + \|\kappa u\|_{L^2(\Omega)}^2\\ &\overset{\forall\varepsilon>0}{\leq} C_\varepsilon\|\kappa\|_{W^{1,\infty}(\Omega)}^2\|u\|_{L^2(\mathcal{B}^\delta)}^2 + \varepsilon\|\kappa\nabla u\|_{L^2(\Omega)}^2.\end{aligned}$$

Finally, since the parameter $\varepsilon > 0$ from Young's inequality can be chosen arbitrarily small, we can absorb the last summand of the right-hand side in the left-hand side of the overall inequality. We end up with

$$|u|_{H^1(\mathcal{B})} \overset{\kappa|_{\mathcal{B}}\equiv 1}{\leq} \|\kappa\nabla u\|_{L^2(\Omega)} \lesssim \|\kappa\|_{W^{1,\infty}(\Omega)}\|u\|_{L^2(\mathcal{B}^\delta)} \overset{\text{Lem. 3.18}}{\lesssim} \frac{1}{\delta}\|u\|_{L^2(\mathcal{B}^\delta)}.$$

This concludes the proof of the discrete Caccioppoli inequality. □

3.8 The single- and multi-step coarsening operators

In this subsection, we do the actual work in the construction of the subspace $V_{\mathcal{B},\mathcal{D},L} \subseteq \mathbb{S}_0^{p,1}(\mathcal{T})$ from Sect. 3.1. We design the so called *single-* and *multi-step coarsening operators*. Given $\mathcal{B} \subseteq \mathcal{T}$, $\delta > 0$ and $u \in \mathbb{S}_{\text{harm}}(\mathcal{B}^\delta)$, the single-step coarsening operator $Q_{\mathcal{B}}^\delta$ produces a "coarse" approximation $Q_{\mathcal{B}}^\delta u \in \mathbb{S}_{\text{harm}}(\mathcal{B})$ with an error $\|u - Q_{\mathcal{B}}^\delta u\|_{L^2(\mathcal{B})} \leq 2^{-1}\|u\|_{L^2(\mathcal{B}^\delta)}$. The prefactor $2^{-1} \in (0,1)$ is essential, as it produces an exponential factor 2^{-L} when $L \in \mathbb{N}$ single-step coarsening operators are combined in a specific manner. This is precisely the idea behind the multi-step coarsening operator $Q_{\mathcal{B}}^{\delta,L}$. Given a function $u \in \mathbb{S}_{\text{harm}}(\mathcal{B}^{\delta L})$, it produces a "coarse" approximation $Q_{\mathcal{B}}^{\delta,L}u \in \mathbb{S}_{\text{harm}}(\mathcal{B})$ with an error $\|u - Q_{\mathcal{B}}^{\delta,L}u\|_{L^2(\mathcal{B})} \leq 2^{-L}\|u\|_{L^2(\mathcal{B}^{\delta L})}$.

As our construction of the single-step coarsening operator in Theorem 3.31 is quite technical, we would like to reveal the underlying ideas first: Assume for a moment that $\mathcal{T}$ is uniform, i.e., $h_{\mathcal{T}} \approx h_{\min,\mathcal{T}}$. Then, a function $u \in \mathbb{S}_{\text{harm}}(\mathcal{B}^\delta)$ is described by up to $\dim \mathbb{S}^{p,0}(\mathcal{T}) \approx \#\mathcal{T} \approx h_{\mathcal{T}}^{-d}$ degrees of freedom. In order to reduce this number, we could approximate $u \approx \Pi_{\mathcal{S}}^p u \in \mathbb{S}^{p,0}(\mathcal{S})$, where $\mathcal{S} \subseteq \text{Pow}(\Omega)$ is a second uniform mesh and where $\Pi_{\mathcal{S}}^p : L^2(\Omega) \longrightarrow \mathbb{S}^{p,0}(\mathcal{S})$ is some kind of approximation operator. As long as $\mathcal{S}$ is coarser than $\mathcal{T}$, i.e. $h_{\mathcal{S}} \gtrsim h_{\mathcal{T}}$, this provides a reduction of the dimension. On the other hand, the typical error bound $\|u - \Pi_{\mathcal{S}}^p u\|_{L^2(\Omega)} \lesssim H|u|_{H^1(\Omega)}$ involves an H^1-norm on the right-hand side. In order to get rid of the H^1-norm, we want to apply the discrete Caccioppoli inequality, Lemma 3.27. For this to work, however, we first need to reduce the global quantity $H|u|_{H^1(\Omega)}$ to the local quantity $H|u|_{H^1(\mathcal{B})}$. This can be done using the discrete cut-off operator $K_{\mathcal{B}}^\delta$ from Definition 3.23. Finally,

the combined operator $\Pi^p_{\mathcal{S}} K^\delta_{\mathcal{B}} : \mathbb{S}_{\mathrm{harm}}(\mathcal{B}^\delta) \longrightarrow \mathbb{S}^{p,0}(\mathcal{S})$ only lacks one more thing: It does not necessarily map into the space $\mathbb{S}_{\mathrm{harm}}(\mathcal{B})$, which is a critical requirement, because we want to iterate the argument by plugging the remainder $\tilde{u} := u - Q^\delta_{\mathcal{B}} u$ of one single-step coarsening operator into another one. Thankfully, we can simply append the orthogonal projection $P_{\mathcal{B}} : L^2(\Omega) \longrightarrow \mathbb{S}_{\mathrm{harm}}(\mathcal{B})$ without losing any of the aforementioned properties.

In the next lemma we provide a construction for the second, coarser mesh $\mathcal{S} \subseteq \mathrm{Pow}(\Omega)$:

Lemma 3.28 *Let $\mathcal{S}_0 \subseteq \mathrm{Pow}(\Omega)$ be an arbitrary mesh and $(\mathcal{S}_\ell)_{\ell \in \mathbb{N}_0}$ be the corresponding sequence of uniform refinements. For every $H > 0$, there exists an $\mathcal{S} \in (\mathcal{S}_\ell)_{\ell \in \mathbb{N}_0}$ with $\sigma_{\mathrm{shp}}(\mathcal{S}) = C(\mathcal{S}_0)$ and $C(\mathcal{S}_0)H \leq h_{\min,\mathcal{S}} \leq h_{\mathcal{S}} \leq H$. In particular, $\mathcal{S}$ is uniform in the sense of Definition 3.2.*

Proof There hold the relations $h_{\mathcal{S}_\ell} = 2^{-\ell} h_{\mathcal{S}_0}$ and $h_{\min,\mathcal{S}_\ell} = 2^{-\ell} h_{\min,\mathcal{S}_0}$. For any given $H > 0$, we choose the mesh $\mathcal{S} := \mathcal{S}_L$, where $L \in \mathbb{N}_0$ is the minimal level satisfying $h_{\mathcal{S}_L} \leq H$. In particular, there also holds the lower bound $H < h_{\mathcal{S}_{L-1}} = 2^{-(L-1)} h_{\mathcal{S}_0} = 2 h_{\mathcal{S}_0} h^{-1}_{\min,\mathcal{S}_0} h_{\min,\mathcal{S}_L} = C(\mathcal{S}_0) h_{\min,\mathcal{S}_L}$. □

The additional mesh $\mathcal{S} \subseteq \mathrm{Pow}(\Omega)$ does not need to be aligned with the original mesh $\mathcal{T} \subseteq \mathrm{Pow}(\Omega)$ at all. The output of the cut-off operator $K^\epsilon_{\mathcal{B}}$ is just an element of $\mathbb{S}^{p,1}_0(\mathcal{T}) \subseteq H^1(\Omega)$, so we need an operator $\Pi_{\mathcal{S}} : H^1(\Omega) \longrightarrow \mathbb{S}^{q,0}(\mathcal{S})$ for some $q \geq 0$. Also, in the case $\mathcal{S} = \mathcal{T}$ the operator should act like a projection on functions from $\mathbb{S}^{p,1}_0(\mathcal{T})$. The simplest solution for these demands is the *$L^2(\Omega)$-orthogonal projection*.

Definition 3.29 We denote by $\Pi^p_{\mathcal{S}} : L^2(\Omega) \longrightarrow \mathbb{S}^{p,0}(\mathcal{S})$ the (global) orthogonal projection from $L^2(\Omega)$ onto the closed subspace $\mathbb{S}^{p,0}(\mathcal{S})$.

In fact, $\Pi^p_{\mathcal{S}}$ coincides with the piecewise L^2-orthogonal projection on the mesh $\mathcal{S}$. This results in desirable *local* properties and bounds.

Lemma 3.30 *The linear operator $\Pi^p_{\mathcal{S}}$ has a local projection property: For every cluster $\mathcal{B} \subseteq \mathcal{S}$ and every function $v \in L^2(\Omega)$ with $v \in \mathbb{S}^{p,0}(\mathcal{B})$, there holds $(\Pi^p_{\mathcal{S}} v)|_{\mathcal{B}} = v|_{\mathcal{B}}$. Furthermore, $\Pi^p_{\mathcal{S}}$ preserves supports: For every $v \in L^2(\Omega)$, we have $\mathrm{supp}_{\mathcal{S}}(\Pi^p_{\mathcal{S}} v) \subseteq \mathrm{supp}_{\mathcal{S}}(v)$. Finally, for every $k \in \{0, \dots, p+1\}$, there hold the stability and error estimates*

$$\forall v \in H^k_{\mathrm{pw}}(\mathcal{S}) : \forall S \in \mathcal{S} : \qquad \sum_{\ell=0}^{k} h_S^\ell |\Pi^p_{\mathcal{S}} v|_{H^\ell(S)} \lesssim \sum_{\ell=0}^{k} h_S^\ell |v|_{H^\ell(S)},$$
$$\sum_{\ell=0}^{k} h_S^\ell |(\mathrm{id} - \Pi^p_{\mathcal{S}})(v)|_{H^\ell(S)} \lesssim h_S^k |v|_{H^k(S)}.$$

Now, we have all the ingredients for the construction of the *single-step coarsening operator*.

Theorem 3.31 *Let $\mathcal{T} \subseteq \mathrm{Pow}(\Omega)$ be a mesh of locally bounded cardinality. Furthermore, let $\mathcal{B} \subseteq \mathcal{T}$ and $\delta > 0$ with $\delta \lesssim 1$. Then, there exists a linear single-step coarsening operator*

$$Q_{\mathcal{B}}^{\delta} : \mathbb{S}_{\mathrm{harm}}(\mathcal{B}^{\delta}) \longrightarrow \mathbb{S}_{\mathrm{harm}}(\mathcal{B})$$

of rank

$$\mathrm{rank}(Q_{\mathcal{B}}^{\delta}) \lesssim \left(1 + \frac{\mathrm{diam}_{\mathcal{T}}(\mathcal{B})}{\delta}\right)^{d\sigma_{\mathrm{card}}}$$

that satisfies the following approximation property: For every $u \in \mathbb{S}_{\mathrm{harm}}(\mathcal{B}^{\delta})$,

$$\|u - Q_{\mathcal{B}}^{\delta} u\|_{L^2(\mathcal{B})} \leq \frac{1}{2}\|u\|_{L^2(\mathcal{B}^{\delta})}.$$

Proof Let $\mathcal{B} \subseteq \mathcal{T}$ and $\delta > 0$ with $\delta \lesssim 1$. For the construction of $Q_{\mathcal{B}}^{\delta}$ we need three operators: First, we use the discrete cut-off operator $K_{\mathcal{B}}^{\varepsilon} : \mathbb{S}^{p,1}(\mathcal{T}) \longrightarrow \mathbb{S}^{p,1}(\mathcal{T})$ from Definition 3.23 with some carefully chosen parameter $\varepsilon > 0$. Second, we apply the piecewise orthogonal projection $\Pi_{\mathcal{S}}^{p} : L^2(\Omega) \longrightarrow \mathbb{S}^{p,0}(\mathcal{S})$ from Definition 3.29 on some suitable mesh $\mathcal{S} \subseteq \mathrm{Pow}(\Omega)$. Third, the result is mapped back into the space $\mathbb{S}_{\mathrm{harm}}(\mathcal{B})$ via the orthogonal projection $P_{\mathcal{B}} : L^2(\Omega) \longrightarrow \mathbb{S}_{\mathrm{harm}}(\mathcal{B})$.

For the precise choice of ε and $\mathcal{S}$ we have to distinguish between two cases: In the more involved case $\delta \geq 20\sigma_{\mathrm{shp}}^{7} h_{\mathcal{B}}$ we choose $\varepsilon := \delta/(5\sigma_{\mathrm{shp}}^{4}) \geq 4\sigma_{\mathrm{shp}}^{3} h_{\mathcal{B}}$ and use the uniform mesh $\mathcal{S} \subseteq \mathrm{Pow}(\Omega)$ from Lemma 3.28 with $h_{\mathcal{S}} \eqsim h_{\min,\mathcal{S}} \eqsim H$, where the parameter $H > 0$ will be specified during the proof. In the degenerate case $\delta < 20\sigma_{\mathrm{shp}}^{7} h_{\mathcal{B}}$ we set $\varepsilon := 4\sigma_{\mathrm{shp}}^{3} h_{\mathcal{B}}$ and use the mesh $\mathcal{S} := \mathcal{T}$ itself.

We define the asserted operator as

$$Q_{\mathcal{B}}^{\delta} := P_{\mathcal{B}} \Pi_{\mathcal{S}}^{p} K_{\mathcal{B}}^{\varepsilon} : \mathbb{S}_{\mathrm{harm}}(\mathcal{B}^{\delta}) \longrightarrow \mathbb{S}_{\mathrm{harm}}(\mathcal{B}).$$

The case $\delta \geq 20\sigma_{\mathrm{shp}}^{7} h_{\mathcal{B}}$: Let $u \in \mathbb{S}_{\mathrm{harm}}(\mathcal{B}^{\delta})$. By Lemma 3.15 we have $h_{\mathcal{B}^{\varepsilon}} \leq \max\{h_{\mathcal{B}}, \sigma_{\mathrm{shp}}\varepsilon\}$, and the assumption on δ implies $h_{\mathcal{B}} \leq \delta/(20\sigma_{\mathrm{shp}}^{7}) = \varepsilon/(4\sigma_{\mathrm{shp}}^{3}) < \varepsilon$, so the maximum in the previous estimate is attained at $\sigma_{\mathrm{shp}}\varepsilon$. Therefore, the parameter $\alpha := 4\sigma_{\mathrm{shp}}^{4}\varepsilon$ satisfies $4\sigma_{\mathrm{shp}}^{3} h_{\mathcal{B}^{\varepsilon}} \leq \alpha \lesssim 1$. In particular, we can apply the discrete Caccioppoli inequality to the set $\mathcal{B}^{\varepsilon}$ and the parameter α. Since $\delta \eqsim \alpha$, this gives the stability estimate for the cut-off operator $K_{\mathcal{B}}^{\varepsilon}$

$$\sum_{\ell=0}^{1} \delta^{\ell} |K_{\mathcal{B}}^{\varepsilon} u|_{H^{\ell}(\Omega)} \overset{\text{Lem. 3.24}}{\lesssim} \sum_{\ell=0}^{1} \alpha^{\ell} |u|_{H^{\ell}(\mathcal{B}^{\varepsilon})} \overset{\text{Lem. 3.27}}{\lesssim} \|u\|_{L^2(\mathcal{B}^{\varepsilon+\alpha})} \overset{\varepsilon+\alpha \leq \delta}{\leq} \|u\|_{L^2(\mathcal{B}^{\delta})}.$$

From Lemmas 3.26 and 3.24 we know that $K_{\mathcal{B}}^{\varepsilon} u \in \mathbb{S}_{\mathrm{harm}}(\mathcal{B})$, hence $P_{\mathcal{B}} K_{\mathcal{B}}^{\varepsilon} u = K_{\mathcal{B}}^{\varepsilon} u$. We conclude $u|_{\mathcal{B}} = (K_{\mathcal{B}}^{\varepsilon} u)|_{\mathcal{B}} = (P_{\mathcal{B}} K_{\mathcal{B}}^{\varepsilon} u)|_{\mathcal{B}}$ and thus

$$\begin{aligned}\|u - Q_{\mathcal{B}}^{\delta} u\|_{L^2(\mathcal{B})} &= \|P_{\mathcal{B}} K_{\mathcal{B}}^{\varepsilon} u - P_{\mathcal{B}} \Pi_{\mathcal{S}}^{p} K_{\mathcal{B}}^{\varepsilon} u\|_{L^2(\mathcal{B})} \underset{\text{Lem. 3.30}}{\leq} \|P_{\mathcal{B}}(\mathrm{id} - \Pi_{\mathcal{S}}^{p})(K_{\mathcal{B}}^{\varepsilon} u)\|_{L^2(\Omega)} \\ &\leq \|(\mathrm{id} - \Pi_{\mathcal{S}}^{p})(K_{\mathcal{B}}^{\varepsilon} u)\|_{L^2(\Omega)} \lesssim H |K_{\mathcal{B}}^{\varepsilon} u|_{H^1(\Omega)} \\ &\lesssim \tfrac{H}{\delta} \|u\|_{L^2(\mathcal{B}^{\delta})}.\end{aligned}$$

In particular, we can choose $H \eqsim \delta > 0$ small enough to establish the asserted error bound.

The case $\delta < 20\sigma_{\mathrm{shp}}^{7} h_{\mathcal{B}}$: Again let $u \in \mathbb{S}_{\mathrm{harm}}(\mathcal{B}^{\delta})$. Exploiting $\mathcal{S} = \mathcal{T}$ and Lemma 3.24, the operator $Q_{\mathcal{B}}^{\delta}$ reduces to $Q_{\mathcal{B}}^{\delta} u = P_{\mathcal{B}} \Pi_{\mathcal{T}}^{p} K_{\mathcal{B}}^{\varepsilon} u = P_{\mathcal{B}} K_{\mathcal{B}}^{\varepsilon} u = K_{\mathcal{B}}^{\varepsilon} u$. Consequently, the error bound becomes trivial:

$$\|u - Q_{\mathcal{B}}^{\delta} u\|_{L^2(\mathcal{B})} = \|u - K_{\mathcal{B}}^{\varepsilon} u\|_{L^2(\mathcal{B})} = \|u - u\|_{L^2(\mathcal{B})} = 0.$$

To find a good upper bound for the rank of $Q_{\mathcal{B}}^{\delta}$, the locally bounded cardinality of $\mathcal{S}$ is crucial. In the case $\delta \geq 20\sigma_{\mathrm{shp}}^{7} h_{\mathcal{B}}$ the mesh $\mathcal{S}$ is uniform and thus of locally bounded cardinality (cf. Lemma 3.3). In the case $\delta < 20\sigma_{\mathrm{shp}}^{7} h_{\mathcal{B}}$ we chose $\mathcal{S} = \mathcal{T}$, which has locally bounded cardinality by assumption.

Next, we abbreviate $B := \bigcup \mathcal{B}^{\varepsilon} \subseteq \mathbb{R}^d$ and compute a common lower bound for $h_{\mathcal{S}(B)}$: In the case $\delta \geq 20\sigma_{\mathrm{shp}}^{7} h_{\mathcal{B}}$ we have $h_{\mathcal{B}} \lesssim \delta$ and $\varepsilon \lesssim \delta$ and $H \eqsim \delta$ by our choice of the parameters. With Lemma 3.28, this implies $h_{\mathcal{B}} + \varepsilon + \delta \lesssim \delta \eqsim H \eqsim h_{\min,\mathcal{S}} \leq h_{\mathcal{S}(B)}$. In the case $\delta < 20\sigma_{\mathrm{shp}}^{7} h_{\mathcal{B}}$, using $\varepsilon \lesssim \delta$ we get in a similar way $h_{\mathcal{B}} + \varepsilon + \delta \lesssim h_{\mathcal{B}} \leq h_{\mathcal{T}(\mathcal{B}^{\varepsilon})} = h_{\mathcal{S}(B)}$ as well.

For every $u \in \mathbb{S}_{\mathrm{harm}}(\mathcal{B}^{\delta})$ we know from Lemmas 3.30 and 3.24 that $\mathrm{supp}_{\mathcal{S}}(\Pi_{\mathcal{S}}^{p} K_{\mathcal{B}}^{\varepsilon} u) \subseteq \mathrm{supp}_{\mathcal{S}}(K_{\mathcal{B}}^{\varepsilon} u) \subseteq \mathcal{S}(B)$. This results in the estimate

$$\begin{aligned}\mathrm{rank}(Q_{\mathcal{B}}^{\delta}) &\leq \dim\{v \in \mathbb{S}^{p,0}(\mathcal{S}) \mid \mathrm{supp}_{\mathcal{S}}(v) \subseteq \mathcal{S}(B)\} \eqsim \#\mathcal{S}(B) \\ &\underset{\text{Def. 2.4}}{\lesssim} (1 + h_{\mathcal{S}(B)}^{-1} \mathrm{diam}_{\mathcal{S}}(\mathcal{S}(B)))^{d\sigma_{\mathrm{card}}} \underset{\text{Lem. 3.15}}{\lesssim} (1 + h_{\mathcal{S}(B)}^{-1}(\mathrm{diam}_{\mathcal{T}}(\mathcal{B}) + h_{\mathcal{B}} + \varepsilon))^{d\sigma_{\mathrm{card}}} \\ &\underset{h_{\mathcal{B}} + \varepsilon \lesssim h_{\mathcal{S}(B)}}{\lesssim} (1 + h_{\mathcal{S}(B)}^{-1} \mathrm{diam}_{\mathcal{T}}(\mathcal{B}))^{d\sigma_{\mathrm{card}}} \underset{\delta \lesssim h_{\mathcal{S}(B)}}{\lesssim} (1 + \delta^{-1} \mathrm{diam}_{\mathcal{T}}(\mathcal{B}))^{d\sigma_{\mathrm{card}}},\end{aligned}$$

which finishes the proof. □

With the single-step coarsening operator at hand, we can iterate to obtain exponential convergence.

Theorem 3.32 *Let $\mathcal{T} \subseteq \mathrm{Pow}(\Omega)$ be a mesh of locally bounded cardinality. Furthermore, let $\mathcal{B} \subseteq \mathcal{T}$ and $\delta > 0$ with $\delta \lesssim 1$. Then, for every $L \in \mathbb{N}$, there exists a linear multi-step coarsening operator*

$$Q_{\mathcal{B}}^{\delta,L} : \mathbb{S}_{\mathrm{harm}}(\mathcal{B}^{\delta L}) \longrightarrow \mathbb{S}_{\mathrm{harm}}(\mathcal{B})$$

of rank

$$\operatorname{rank}(Q_{\mathcal{B}}^{\delta,L}) \lesssim \left(L + \frac{\operatorname{diam}_{\mathcal{T}}(\mathcal{B})}{\delta}\right)^{d\sigma_{\mathrm{card}}+1}$$

that satisfies the following approximation property: For every $u \in \mathbb{S}_{\mathrm{harm}}(\mathcal{B}^{\delta L})$*, there holds*

$$\|u - Q_{\mathcal{B}}^{\delta,L}u\|_{L^2(\mathcal{B})} \le 2^{-L}\|u\|_{L^2(\mathcal{B}^{\delta L})}.$$

Proof Let $\mathcal{B} \subseteq \mathcal{T}$ and $\delta > 0$ with $\delta \lesssim 1$ as well as $L \in \mathbb{N}$. We define a sequence of nested element sets $\mathcal{B} \subseteq \mathcal{B}_0 \subseteq \cdots \subseteq \mathcal{B}_L \subseteq \mathcal{B}^{\delta L}$ inductively by $\mathcal{B}_0 := \mathcal{B}$ and $\mathcal{B}_{\ell+1} := (\mathcal{B}_\ell)^\delta$. Using the corresponding single-step coarsening operators $Q_\ell := Q_{\mathcal{B}_\ell}^{\delta} : \mathbb{S}_{\mathrm{harm}}(\mathcal{B}_{\ell+1}) \longrightarrow \mathbb{S}_{\mathrm{harm}}(\mathcal{B}_\ell)$ from Theorem 3.31, we make the following definition:

$$\forall u \in \mathbb{S}_{\mathrm{harm}}(\mathcal{B}^{\delta L}) : \qquad Q_{\mathcal{B}}^{\delta,L}u := u - (\mathrm{id} - Q_0)\circ\cdots\circ(\mathrm{id} - Q_{L-1})(u) \in \mathbb{S}_{\mathrm{harm}}(\mathcal{B}).$$

Using the alternative representation

$$Q_{\mathcal{B}}^{\delta,L}u = -\sum_{\pi\in\{0,1\}^L\setminus\{0\}} (-Q_0)^{(\pi_0)}\circ\cdots\circ(-Q_{L-1})^{(\pi_{L-1})}(u),$$

we infer

$$\begin{aligned}
\operatorname{rank}(Q_{\mathcal{B}}^{\delta,L}) &\le \sum_{\ell=0}^{L-1}\operatorname{rank}(Q_\ell) \overset{\text{Thm. 3.31}}{\lesssim} \sum_{\ell=0}^{L-1}(1+\delta^{-1}\operatorname{diam}_{\mathcal{T}}(\mathcal{B}_\ell))^{d\sigma_{\mathrm{card}}} \\
&\overset{\text{Lem. 3.15}}{\lesssim} \sum_{\ell=0}^{L-1}(1+\ell+\delta^{-1}\operatorname{diam}_{\mathcal{T}}(\mathcal{B}))^{d\sigma_{\mathrm{card}}} \le L(L+\delta^{-1}\operatorname{diam}_{\mathcal{T}}(\mathcal{B}))^{d\sigma_{\mathrm{card}}} \\
&\le (L+\delta^{-1}\operatorname{diam}_{\mathcal{T}}(\mathcal{B}))^{d\sigma_{\mathrm{card}}+1}.
\end{aligned}$$

Finally, the definition of $Q_{\mathcal{B}}^{\delta,L}$ was such that the error bound becomes elementary: For every $u \in \mathbb{S}_{\mathrm{harm}}(\mathcal{B}^{\delta L})$, iteration of Theorem 3.31 gives

$$\|u - Q_{\mathcal{B}}^{\delta,L}u\|_{L^2(\mathcal{B})} = \|(\mathrm{id} - Q_0)\circ\cdots\circ(\mathrm{id} - Q_{L-1})(u)\|_{L^2(\mathcal{B}_0)} \le 2^{-L}\|u\|_{L^2(\mathcal{B}^{\delta L})},$$

which finishes the proof. □

3.9 Putting everything together

We can finally answer the question of how to find the subspace $V_{\mathcal{B},\mathcal{D},L} \subseteq L^2(\Omega)$ from Sect. 3.1. After that, the Proof of Theorem 2.15 is just a matter of putting everything together.

Theorem 3.33 *Let $\mathcal{T} \subseteq \mathrm{Pow}(\Omega)$ be a mesh of locally bounded cardinality and $\mathcal{B}$, $\mathcal{D} \subseteq \mathcal{T}$ clusters satisfying*

$$0 < \mathrm{diam}_{\mathcal{T}}(\mathcal{B}) \leq \sigma_{\mathrm{adm}} \mathrm{dist}_{\mathcal{T}}(\mathcal{B}, \mathcal{D}).$$

Then, for every $L \in \mathbb{N}$, there exists a subspace

$$V_{\mathcal{B},\mathcal{D},L} \subseteq \mathbb{S}_0^{p,1}(\mathcal{T})$$

of dimension

$$\dim V_{\mathcal{B},\mathcal{D},L} \lesssim L^{d\sigma_{\mathrm{card}}+1}$$

that satisfies the following approximation property: For every $f \in L^2(\Omega)$ with $\mathrm{supp}_{\mathcal{T}}(f) \subseteq \mathcal{D}$ there holds

$$\inf_{v \in V_{\mathcal{B},\mathcal{D},L}} \|S_{\mathcal{T}} f - v\|_{L^2(\mathcal{B})} \lesssim 2^{-L} \|f\|_{L^2(\mathcal{D})}.$$

Proof Let $\mathcal{B}, \mathcal{D} \subseteq \mathcal{T}$ with $0 < \mathrm{diam}_{\mathcal{T}}(\mathcal{B}) \leq \sigma_{\mathrm{adm}} \mathrm{dist}_{\mathcal{T}}(\mathcal{B}, \mathcal{D})$. For every given $L \in \mathbb{N}$, we make the choice $\delta := \mathrm{diam}_{\mathcal{T}}(\mathcal{B})/(2\sigma_{\mathrm{adm}} L) > 0$ and use the space

$$V_{\mathcal{B},\mathcal{D},L} := \mathrm{ran}(Q_{\mathcal{B}}^{\delta,L}) \subseteq \mathbb{S}_0^{p,1}(\mathcal{T}).$$

Here, $Q_{\mathcal{B}}^{\delta,L} : \mathbb{S}_{\mathrm{harm}}(\mathcal{B}^{\delta L}) \longrightarrow \mathbb{S}_{\mathrm{harm}}(\mathcal{B})$ is the multi-step coarsening operator from Theorem 3.32. Using Theorem 3.32 and the definition of δ, we can bound the dimension by

$$\dim V_{\mathcal{B},\mathcal{D},L} = \mathrm{rank}(Q_{\mathcal{B}}^{\delta,L}) \lesssim \left(L + \frac{\mathrm{diam}_{\mathcal{T}}(\mathcal{B})}{\delta} \right)^{d\sigma_{\mathrm{card}}+1} \lesssim L^{d\sigma_{\mathrm{card}}+1}.$$

To see the approximation properties, let $f \in L^2(\Omega)$ with $\mathrm{supp}_{\mathcal{T}}(f) \subseteq \mathcal{D}$. By the definition of $\mathcal{B}^{\delta L}$ and $\mathrm{dist}_{\mathcal{T}}(\mathcal{B}^{\delta L}, \mathcal{D})$, there exist elements $B \in \mathcal{B}$, $C \in \mathcal{B}^{\delta L}$, $D \in \mathcal{D}$ such that $\mathrm{dist}_{\mathcal{T}}(B, C) \leq \delta L$ and $\mathrm{dist}_{\mathcal{T}}(\mathcal{B}^{\delta L}, \mathcal{D}) = \mathrm{dist}_{\mathcal{T}}(C, D)$. Using the triangle inequality of the mesh metric $\mathrm{dist}_{\mathcal{T}}(\cdot, \cdot)$, we conclude $\mathrm{dist}_{\mathcal{T}}(\mathcal{B}, \mathcal{D}) \leq \mathrm{dist}_{\mathcal{T}}(B, D) \leq \mathrm{dist}_{\mathcal{T}}(B, C) + \mathrm{dist}_{\mathcal{T}}(C, D) \leq \delta L + \mathrm{dist}_{\mathcal{T}}(\mathcal{B}^{\delta L}, \mathcal{D})$. Now, exploiting the definition of δ and the assumptions on $\mathcal{B}, \mathcal{D}$, we obtain

$$\mathrm{dist}_{\mathcal{T}}(\mathcal{B}^{\delta L}, \mathcal{D}) \geq \mathrm{dist}_{\mathcal{T}}(\mathcal{B}, \mathcal{D}) - \delta L = \mathrm{dist}_{\mathcal{T}}(\mathcal{B}, \mathcal{D}) - \frac{\mathrm{diam}_{\mathcal{T}}(\mathcal{B})}{2\sigma_{\mathrm{adm}}} \geq \frac{\mathrm{diam}_{\mathcal{T}}(\mathcal{B})}{2\sigma_{\mathrm{adm}}} > 0.$$

Then, Lemma 3.26 implies $S_{\mathcal{T}} f \in \mathbb{S}_{\mathrm{harm}}(\mathcal{B}^{\delta L})$ and ultimately

$$\inf_{v \in V_{\mathcal{B},\mathcal{D},L}} \|S_{\mathcal{T}} f - v\|_{L^2(\mathcal{B})} \leq \|S_{\mathcal{T}} f - Q_{\mathcal{B}}^{\delta,L}(S_{\mathcal{T}} f)\|_{L^2(\mathcal{B})} \overset{\text{Thm. 3.31}}{\leq} 2^{-L} \|S_{\mathcal{T}} f\|_{L^2(\mathcal{B}^{\delta L})} \overset{\text{Def. 3.15}}{\lesssim} 2^{-L} \|f\|_{L^2(\mathcal{D})},$$

which finishes the proof. □

We close this section with the Proof of Theorem 2.15.

Proof (of Theorem 2.15) Let $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ be the matrix from Definition 2.9 and $r \in \mathbb{N}$ a given block rank bound. We define the asserted $\mathcal{H}$-matrix approximant $\boldsymbol{B} \in \mathbb{R}^{N\times N}$ to $\boldsymbol{A}^{-1}$ in a block-wise fashion:

First, for every admissible block $(I, J) \in \mathbb{P}_{\text{adm}}$, we denote the corresponding index patches by $\mathcal{B} := \mathcal{T}(I) \subseteq \mathcal{T}$ and $\mathcal{D} := \mathcal{T}(J) \subseteq \mathcal{T}$. From Lemma 2.12 we know that $0 < \operatorname{diam}_{\mathcal{T}}(\mathcal{B}) \leq \sigma_{\text{adm}} \operatorname{dist}_{\mathcal{T}}(\mathcal{B}, \mathcal{D})$. Furthermore, let $C > 0$ be the constant from the dimension bound in Theorem 3.33. We set $\sigma_{\text{exp}} := (1/C)^{1/(d\sigma_{\text{card}}+1)} \ln(2) > 0$ and $L := \lfloor (r/C)^{1/(d\sigma_{\text{card}}+1)} \rfloor \in \mathbb{N}$. Then, Theorem 3.33 provides a subspace $V_{\mathcal{B},\mathcal{D},L} \subseteq \mathbb{S}_0^{p,1}(\mathcal{T}) \subseteq L^2(\Omega)$. We apply Lemma 3.13 to the subspace $V_{\mathcal{B},\mathcal{D},L} \subseteq L^2(\Omega)$ and get matrices $\boldsymbol{X}_{I,J}^r \in \mathbb{R}^{I\times\tilde{r}}$ and $\boldsymbol{Y}_{I,J}^r \in \mathbb{R}^{J\times\tilde{r}}$ of size $\tilde{r} \leq \dim V_{\mathcal{B},\mathcal{D},L}$. We set

$$\boldsymbol{B}|_{I\times J} := \boldsymbol{X}_{I,J}^r (\boldsymbol{Y}_{I,J}^r)^T.$$

Second, for every small block $(I, J) \in \mathbb{P}_{\text{small}}$, we make the trivial choice

$$\boldsymbol{B}|_{I\times J} := \boldsymbol{A}^{-1}|_{I\times J}.$$

By Definition 2.13, we have $\boldsymbol{B} \in \mathcal{H}(\mathbb{P}, \tilde{r})$ with a block rank bound

$$\tilde{r} \leq \dim V_{\mathcal{B},\mathcal{D},L} \overset{\text{Thm. 3.33}}{\leq} CL^{d\sigma_{\text{card}}+1} \leq r.$$

For the error we get

$$\begin{aligned}
\|\boldsymbol{A}^{-1} - \boldsymbol{B}\|_2 &\overset{\text{Lem. 3.12}}{\lesssim} \ln(N) \cdot \max_{(I,J)\in\mathbb{P}_{\text{adm}}} \|\boldsymbol{A}^{-1}|_{I\times J} - \boldsymbol{X}_{I,J}^r (\boldsymbol{Y}_{I,J}^r)^T\|_2 \\
&\overset{\text{Lem. 3.13}}{\leq} \ln(N)\|\Lambda\|^2 \cdot \max_{\substack{\mathcal{B},\mathcal{D}\subseteq\mathcal{T} \\ \text{admissible}}} \sup_{\substack{f\in L^2(\Omega): \\ \operatorname{supp}_{\mathcal{T}}(f)\subseteq\mathcal{D}}} \inf_{v\in V_{\mathcal{B},\mathcal{D},L}} \frac{\|S_{\mathcal{T}} f - v\|_{L^2(\mathcal{B})}}{\|f\|_{L^2(\mathcal{D})}} \\
&\overset{\text{Thm. 3.33}}{\lesssim} \ln(N)\|\Lambda\|^2 2^{-L} \\
&\lesssim \ln(N)\|\Lambda\|^2 \exp(-\sigma_{\text{exp}} r^{1/(d\sigma_{\text{card}}+1)}).
\end{aligned}$$

Finally, it only remains to bound the norm of Λ:

$$\|\Lambda\|^2 \overset{\text{Def. 2.6}}{\lesssim} h_{\min,\mathcal{T}}^{-d} \overset{\text{Def. 2.4}}{\lesssim} h_{\mathcal{T}}^{-d\sigma_{\text{card}}} \overset{\text{Lem. 3.1}}{\lesssim} \#\,\mathcal{T}^{\sigma_{\text{card}}} \eqsim (\dim \mathbb{S}_0^{p,1}(\mathcal{T}))^{\sigma_{\text{card}}} = N^{\sigma_{\text{card}}}.$$

This concludes the proof of the main result, Theorem 2.15. □

4 Numerical results

In this subsection, we illustrate the validity of Theorem 2.15 by means of a numerical example: For the geometry we choose the *L-shaped domain* $\Omega := ((0,1) \times (0,1)) \backslash ([1/2,1] \times [1/2,1]) \subseteq \mathbb{R}^2$ in two space dimensions. The PDE coefficients for the model problem from Sect. 2.1 are given by

$$a_1(x) := \begin{pmatrix} 10 & -1 \\ -1 & 1 \end{pmatrix}, \; a_2(x) := \begin{pmatrix} 10x_2 \\ 0 \end{pmatrix}, \; a_3(x) := 1.$$

The mesh $\mathcal{T}$ is *graded* in the sense of Definition 3.4 towards $\Gamma := \{(1/2, 1/2)\}$ with exponent $\alpha := 5$ and the coarse mesh width $H := 0.0095$. We use the spline space $\mathbb{S}_0^{1,1}(\mathcal{T})$ ($p = 1$, globally continuous, piecewise linear) and the well-known basis of *hat-functions* $\{\varphi_1, \dots, \varphi_N\} \subseteq \mathbb{S}_0^{1,1}(\mathcal{T})$. The block partition $\mathbb{P}$ is constructed from a *geometrically balanced cluster tree* $\mathbb{T}_N$ as suggested in [16]. We choose the parameters $\sigma_{\text{adm}} := 2$ and $\sigma_{\text{small}} := 25$ (cf. Definition 2.10). For the rank bound we choose the range $r \in \{1, \dots, 50\}$.

Unfortunately, the $\mathcal{H}$-matrix approximant $\boldsymbol{B} \in \mathbb{R}^{N \times N}$ from our proof is only a theoretical tool and inaccessible for an implementation in a computer system. Hence, we revert to a *block-wise singular value decomposition*: First, we compute the exact inverse $\boldsymbol{A}^{-1} \in \mathbb{R}^{N \times N}$ explicitly. Then, for every admissible block $(I, J) \in \mathbb{P}_{\text{adm}}$, we perform the singular value decomposition $\boldsymbol{A}^{-1}|_{I \times J} = \boldsymbol{U} \boldsymbol{\Sigma} \boldsymbol{V}^T \in \mathbb{R}^{I \times J}$. Here, $\boldsymbol{U} \in \mathbb{R}^{I \times I}$, $\boldsymbol{V} \in \mathbb{R}^{J \times J}$ are orthogonal and $\boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1, \dots, \sigma_{\min\{\#I, \#J\}}) \in \mathbb{R}^{I \times J}$ contains the corresponding singular values $\sigma_1 \geq \cdots \geq \sigma_{\min\{\#I, \#J\}} \geq 0$. Now, for the approximant we use $\boldsymbol{B}|_{I \times J} := \boldsymbol{U}_r \boldsymbol{\Sigma}_r \boldsymbol{V}_r^T \in \mathbb{R}^{I \times J}$, where $\boldsymbol{U}_r \in \mathbb{R}^{I \times r}$, $\boldsymbol{\Sigma}_r \in \mathbb{R}^{r \times r}$ and $\boldsymbol{V}_r \in \mathbb{R}^{J \times r}$ are the first r columns of $\boldsymbol{U}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{V}$, respectively. Recall from the theory of singular value decompositions (e.g., [21]) that

$$\|\boldsymbol{A}^{-1}|_{I \times J} - \boldsymbol{B}|_{I \times J}\|_2 = \min_{\substack{\boldsymbol{C} \in \mathbb{R}^{I \times J}: \\ \operatorname{rank}(\boldsymbol{C}) \leq r}} \|\boldsymbol{A}^{-1}|_{I \times J} - \boldsymbol{C}\|_2 = \sigma_{r+1}.$$

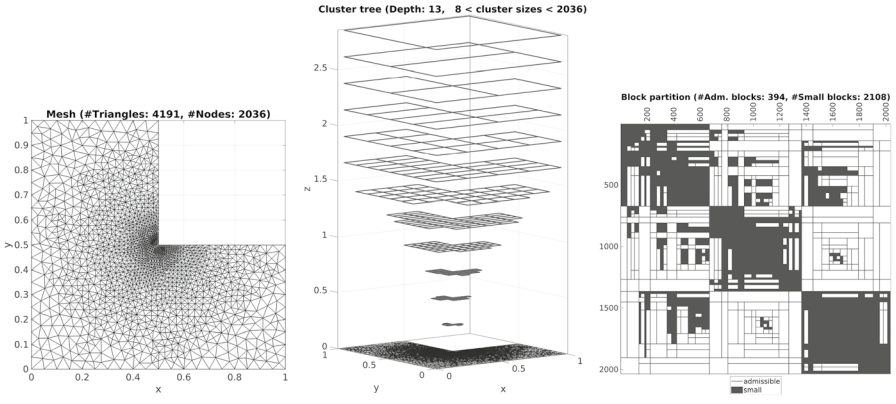

Fig. 1 The mesh $\mathcal{T}$, the cluster tree $\mathbb{T}_N$ and the block partition $\mathbb{P}$ for $N \approx 2.000$ degrees of freedom

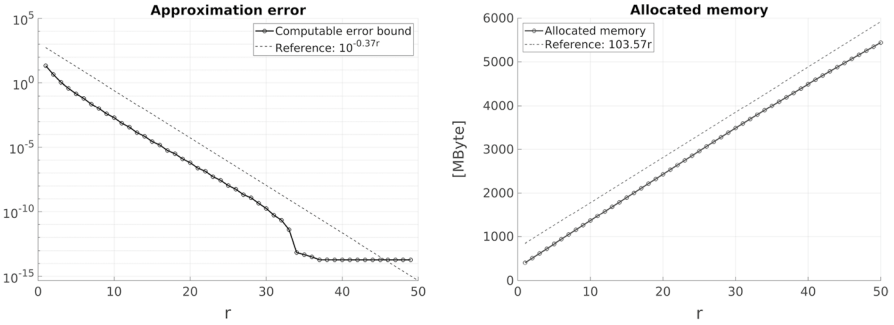

Fig. 2 Approximation error and memory allocation for $N \approx 72.000$ degrees of freedom

In particular, we end up with the following *computable* error bound (cf. [21, Lemma 6.5.8])

$$\begin{aligned}\|\boldsymbol{A}^{-1} - \boldsymbol{B}\|_2 &\lesssim \operatorname{depth}(\mathbb{T}_{N\times N}) \cdot \max_{(I,J)\in\mathbb{P}} \|\boldsymbol{A}^{-1}|_{I\times J} - \boldsymbol{B}|_{I\times J}\|_2 \\ &= \operatorname{depth}(\mathbb{T}_{N\times N}) \cdot \max_{(I,J)\in\mathbb{P}} \sigma_{r+1}(\boldsymbol{A}^{-1}|_{I\times J}).\end{aligned}$$

The numerical example is implemented in MATLAB. For the inversion of the full matrix $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ we use MATLAB's built-in procedure `inv(...)`. For the singular value decompositions we use `svds(...)`. Recall that an exact matrix inversion needs $\mathcal{O}(N^2)$ memory and $\mathcal{O}(N^3)$ time to compute, which effectively restricts the maximal feasible problem size to $N \approx 70.000$ on our machine.

In Fig. 1, we chose $N \approx 2.000$ degrees of freedom. The elements are graded towards the reentrant corner with a grading exponent $\alpha = 5$. The cluster tree $\mathbb{T}_N$ is clearly deeper near the grading center. The block partition $\mathbb{P}$ uses sorted indices internally. Only a few admissible blocks are far away from the diagonal, lots of small blocks agglomerate along the diagonal. The sparsity pattern becomes more pronounced as $N \to \infty$.

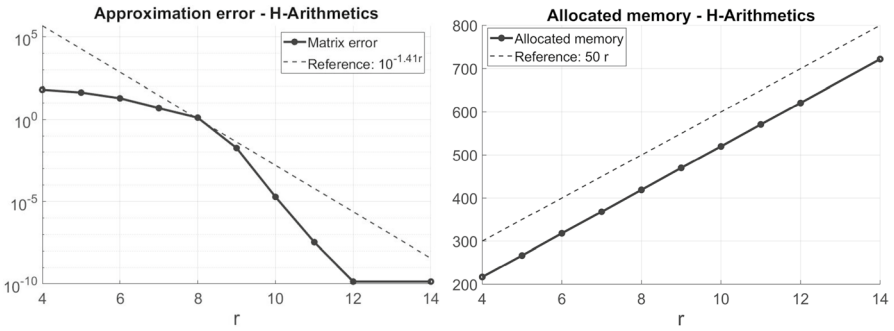

Fig. 3 Approximation error and memory allocation (in MB) for $N \approx 30.000$ degrees of freedom using HLiB

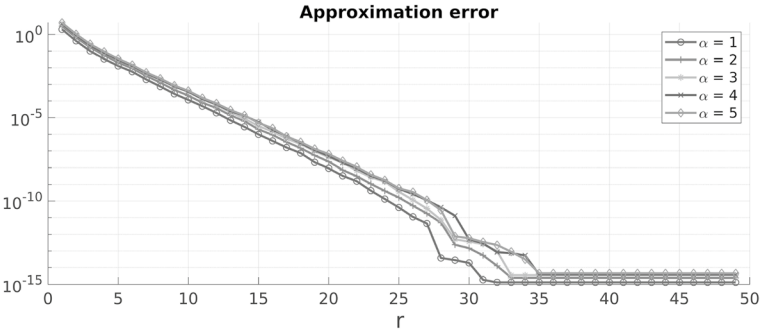

Fig. 4 Comparison of approximation errors for different grading exponents, $\alpha \in \{1, 2, 3, 4, 5\}$. The number of degrees of freedom was kept constant at roughly $N \approx 17.500$ throughout all five runs

In Fig. 2, we chose $N \approx 72.000$ degrees of freedom. The computable error bound from above (for $r \in \{1, \dots, 50\}$) is depicted on a linear abscissa and a logarithmic ordinate. The values are below a straight line with slope -0.37 indicating an *exponential decay* $\text{error}(r) \lesssim 10^{-0.37r}$. This is even better than the asserted bound from Theorem 2.15. The allocated memory in MBytes is plotted on a linear abscissa and a linear ordinate. The values are below a straight line with slope 103.57 indicating a *polynomial growth* $\text{memory}(r) \lesssim r$. Choosing a rank bound $r = 37$, for example, gives an approximation error $\approx 10^{-14}$ and uses ≈ 4.2 GByte memory. The full system matrix takes ≈ 41.4 GByte memory.

In Fig. 3, we chose $N \approx 30.000$ degrees of freedom on a graded mesh with grading exponent $\alpha = 5$ and this time computed the $\mathcal{H}$-matrix approximation using the C-Library HLiB, [4]. The approximation to the inverse matrix is computed using the $\mathcal{H}$-matrix arithmetic of HLiB, specifically the hierarchical *LU*-decomposition. The errors shown are $\|\mathbf{I} - \mathbf{A}(\mathbf{L}_{\mathcal{H}}\mathbf{U}_{\mathcal{H}})^{-1}\|_2$, which is an upper bound for the relative error and computable without computing the inverse matrix. Again, we observe exponential convergence with respect to the rank r and linear growth in the memory requirements.

Finally, in Fig. 4, we chose $N \approx 17.500$ degrees of freedom and multiple grading exponents in the range $\{1, 2, 3, 4, 5\}$. The case $\alpha = 1$ corresponds to a uniform mesh, whereas $\alpha = 5$ is "heavily" graded. Again, the computable error bound from above is shown on a linear abscissa and a logarithmic ordinate. As suggested by our main result, Theorem 2.15, the convergence speed deteriorates as α is increased.

Acknowledgements NA was funded by the Austrian Science Fund (FWF) Project P 28367 and JMM was supported by the Austrian Science Fund (FWF) by the special research program Taming complexity in PDE systems (Grant SFB F65).

Funding Open access funding provided by TU Wien (TUW).

References

1. Angleitner, N., Faustmann, M., Melenk, J.M.: Using $\mathcal{H}$-matrices to approximate inverse radial basis function interpolation matrices, in preparation (2021)
2. Bebendorf, M.: Efficient inversion of Galerkin matrices of general second-order elliptic differential operators with nonsmooth coefficients. Math. Comp. **74**, 1179–1199 (2005)
3. Bebendorf, M.: Why finite element discretizations can be factored by triangular hierarchical matrices. SIAM J. Numer. Anal. **45**(4), 1472–1494 (2007)
4. Börm, S., Grasedyck, L.: H-lib - a library for H- and H^2-matrices, available at http://www.hlib.org (1999)
5. Bebendorf, M., Hackbusch, W.: Existence of $\mathcal{H}$-matrix approximants to the inverse FE-matrix of elliptic operators with L^∞-coefficients. Numer. Math. **95**(1), 1–28 (2003)
6. Babuška, I., Kellogg, R.B., Pitkäranta, J.: Direct and inverse error estimates for finite elements with mesh refinements. Numer. Math. **33**, 447–471 (1979)
7. Börm, S.: Approximation of solution operators of elliptic partial differential equations by $\mathcal{H}$- and $\mathcal{H}^2$ -matrices. Numer. Math. **115**(2), 165–193 (2010)
8. Börm, S.: Efficient Numerical Methods for Non-local Operators, EMS Tracts in Mathematics, vol. 14. European Mathematical Society (EMS), Zürich (2010)
9. Ciarlet, P.G.: The finite element method for elliptic problems, North-Holland Publishing Co., Amsterdam-New York-Oxford, 1978, Studies in Mathematics and its Applications, Vol. **4**
10. Clément, Ph.: Approximation by finite element functions using local regularization. Rev. Française Automat. Informat. Recherche Opérationnelle Sér. **9**(R–2), 77–84 (1975)
11. Faustmann, M., Melenk, J.M., Praetorius, D.: $\mathcal{H}$-matrix approximability of the inverses of FEM matrices. Numer. Math. **131**(4), 615–642 (2015)
12. Faustmann, M., Melenk, J.M., Praetorius, D.: Existence of $\mathcal{H}$-matrix approximants to the inverse of BEM matrices: the simple-layer operator. Math. Comp. **85**, 119–152 (2016)
13. Faustmann, M., Melenk, J.M., Praetorius, D.: Existence of $\mathcal{H}$-matrix approximants to the inverse of BEM matrices: the hyper-singular integral operator. IMA J. Numer. Anal. **37**(3), 1211–1244 (2017)
14. Grasedyck, L., Hackbusch, W.: Construction and arithmetics of $\mathcal{H}$-matrices. Computing **70**(4), 295–334 (2003)
15. Grasedyck, L., Hackbusch, W., Kriemann, R.: Performance of $\mathcal{H}$-LU preconditioning for sparse matrices. Comput. Methods Appl. Math. **8**(4), 336–349 (2008)
16. Grasedyck, L., Hackbusch, W., Le Borne, S.: Adaptive geometrically balanced clustering of $\mathcal{H}$ -matrices. Computing **73**(1), 1–23 (2004)
17. Grasedyck, L., Kriemann, R., Le Borne, S.: Parallel black box $\mathcal{H}$-LU preconditioning for elliptic boundary value problems. Comput. Vis. Sci. **11**(4–6), 273–291 (2008)
18. Greengard, L., Rokhlin, V.: A new version of the fast multipole method for the Laplace in three dimensions. Acta Numerica **1997**, 229–269 (1997)
19. Grasedyck, L.: Theorie und Anwendungen Hierarchischer Matrizen, Ph.D. thesis, Universität Kiel, (2001)
20. Hackbusch, W.: A sparse matrix arithmetic based on $\mathcal{H}$-matrices. Introd. $\mathcal{H}$-Matrices Comput. **62**(2), 89–108 (1999)
21. Hackbusch, W.: Hierarchical Matrices: Algorithms and Analysis, Springer Series in Computational Mathematics, vol. 49. Springer, Heidelberg (2015)
22. Lintner, M.: The eigenvalue problem for the 2D Laplacian in $\mathcal{H}$- matrix arithmetic and application to the heat and wave equation. Computing **72**(3–4), 293–323 (2004)
23. Rokhlin, V.: Rapid solution of integral equations of classical potential theory. J. Comput. Phys. **60**, 187–207 (1985)

24. Schneider, R.: Multiskalen- und Wavelet-Matrixkompression: Analysisbasierte Methoden zur effizienten Lösung großer vollbesetzter Gleichungssysteme. Advances in Numerical Mathematics. Teubner, Stuttgart (1998)
25. Tausch, J., White, J.: Multiscale bases for the sparse representation of boundary integral operators on complex geometry. SIAM J. Sci. Comput. **24**(5), 1610–1629 (2003)
26. von Petersdorff, T., Schwab, Ch., Schneider, R.: Multiwavelets for second-kind integral equations. SIAM J. Numer. Anal. **34**(6), 2212–2227 (1997)
27. Wendland, H.: Scattered Data Approximation, Cambridge Monographs on Applied and Computational Mathematics, vol. 17. Cambridge University Press, Cambridge (2005)