

ASC Report No. 14/2020

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

N. Angleitner, M. Faustmann, J.M. Melenk

Most recent ASC Reports

- 13/2020 *F. Achleitner, C. Kuehn, J.M. Melenk, A. Rieder*
Metastable speeds in the fractional Allen-Cahn equation
- 12/2020 *A. Haberl, D. Praetorius, S. Schimanko, and M. Vohralik*
Convergence and quasi-optimal cost of adaptive algorithms for nonlinear operators including iterative linearization and algebraic solver
- 11/2020 *L. Banjai, J.M. Melenk, and C. Schwab*
hp-FEM for reaction-diffusion equations
II. Robust exponential convergence for multiple length scales in corner domains
- 10/2020 *G. Gantner and D. Praetorius*
Adaptive BEM for elliptic PDE systems,
Part I: Abstract framework for weakly-singular integral equations
- 09/2020 *C. Helmer and A. Jüngel*
Analysis of Maxwell-Stefan systems for heat conducting fluid mixtures
- 08/2020 *L. Mascotto, J.M. Melenk, I. Perugia, A. Rieder*
FEM-BEM mortar coupling for the Helmholtz problem in three dimensions
- 07/2020 *R. Becker, M. Innerberger, and D. Praetorius*
Optimal convergence rates for goal-oriented FEM with quadratic goal functional
- 06/2020 *A. Rieder, F.-J. Sayas, J.M. Melenk*
Runge-Kutta approximation for C_0 -semigroups in the graph norm with applications to time domain boundary integral equations
- 05/2020 *A. Arnold, C. Schmeiser, and B. Signorello*
Propagator norm and sharp decay estimates for Fokker-Planck equations with linear drift
- 04/2020 *G. Gantner, A. Haberl, D. Praetorius, and S. Schimanko*
Rate optimality of adaptive finite element methods with respect to the overall computational costs

Institute for Analysis and Scientific Computing
Vienna University of Technology
Wiedner Hauptstraße 8–10
1040 Wien, Austria

E-Mail: admin@asc.tuwien.ac.at
WWW: <http://www.asc.tuwien.ac.at>
FAX: +43-1-58801-10196

ISBN 978-3-902627-00-1

© Alle Rechte vorbehalten. Nachdruck nur mit Genehmigung des Autors.



APPROXIMATING INVERSE FEM MATRICES ON NON-UNIFORM MESHES WITH \mathcal{H} -MATRICES

NIKLAS ANGLEITNER, MARKUS FAUSTMANN, JENS MARKUS MELENK

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 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.

1. INTRODUCTION

Discretizations of elliptic partial differential equations on a domain $\Omega \subseteq \mathbb{R}^d$ using the classic finite element method (FEM) usually produce sparse linear systems of equations $\mathbf{A}\mathbf{x} = \mathbf{b}$ 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 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 tailored 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 [Hac99] and extensively studied in the monograph [Hac15], provide a different solution approach to this problem that does not suffer from the drawbacks of classic 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, [Rok85, GR97], or wavelet methods, [vPSS97, Sch98, TW03], 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, [Gra01, GH03, Hac15]. Therefore, using an \mathcal{H} -matrix approximation to the inverse \mathbf{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, [Beb07, GHK08, GKL08]. Nonetheless, we mention that the accuracy in terms of the maximal blockwise rank of the computed approximations to \mathbf{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 [Gra01], several works in the literature provide existence results of approximations to the inverse matrices in the \mathcal{H} -matrix format. For the inverses of FEM matrices, e.g., see [BH03, Beb05, Bör10, FMP15] and for inverse BEM matrices, see [FMP16, FMP17]. 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 classic example are graded meshes for the solution of elliptic problems in corner domains, [BKP79].

INSTITUTE FOR ANALYSIS AND SCIENTIFIC COMPUTING (INST. E 101), VIENNA UNIVERSITY OF TECHNOLOGY, WIEDNER HAUPTSTRASSE 8-10, 1040 WIEN, AUSTRIA

E-mail address: niklas.angleitner@tuwien.ac.at, markus.faustmann@tuwien.ac.at, melenk@tuwien.ac.at.

Date: May 15, 2020.

2010 *Mathematics Subject Classification.* Primary: 65F50, Secondary: 65F30, 65N30.

Key words and phrases. FEM, \mathcal{H} -matrices, Approximability, Non-uniform meshes.

In this article, we generalize the results of [FMP15] for quasiuniform meshes to meshes of so called *locally bounded cardinality* (cf. D.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. [GHLB04], 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 [FMP15]. While not repeated in this article, we mention that the (mostly algebraic) techniques of [FMP15, 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 Section 2 we introduce all necessary definitions and concepts and state our main result, T.2.13. Section 3 is dedicated to the proof of the main result. The main technical contribution is the discrete Caccioppoli-type estimate presented in L.3.28, 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, Section 4 provides a numerical example that illustrates our main result.

Concerning notation: We write “ $a \lesssim b$ ” iff 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 *independent* of all critical parameters such as the mesh width. We write $a \approx b$, 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., $\mathbf{A} \in \mathbb{R}^{N \times N}$ and $\mathbf{f} \in \mathbb{R}^N$. For all $x \in \mathbb{R}^d$ and $\varepsilon > 0$, we write $\text{Ball}_2(x, r) := \{y \in \mathbb{R}^d \mid \|y - x\|_2 < \varepsilon\}$ 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 by $W^{k,q}(\Omega)$. For a given mesh \mathcal{T} , we denote by $W_{\text{pw}}^{k,q}(\mathcal{T})$ the broken Sobolev space consisting of elementwise functions from $W^{k,q}$. For all $v \in W_{\text{pw}}^{k,q}(\mathcal{T})$ and $\mathcal{B} \subseteq \mathcal{T}$, we set $|v|_{W^{k,q}(\mathcal{B})} := (\sum_{T \in \mathcal{B}} |v|_{W^{k,q}(T)}^q)^{1/q}$ and $|v|_{W^{k,\infty}(\mathcal{B})} := \max_{T \in \mathcal{B}} |v|_{W^{k,\infty}(T)}$. Similarly, $C_{\text{pw}}^0(\mathcal{T})$ denotes the space of piecewise continuous functions. 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 $\text{supp}_{\mathcal{T}}(v) := \{T \in \mathcal{T} \mid v|_T \neq 0\}$. In particular, we have $\text{supp}_{\mathcal{T}}(v) \subseteq \mathcal{T}$ and $\bigcup \text{supp}_{\mathcal{T}}(v) \subseteq \mathbb{R}^d$, which slightly differs from the usual definition of a support, namely, $\text{supp}(v) := \overline{\{x \in \Omega \mid v(x) \neq 0\}} \subseteq \mathbb{R}^d$.

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 $u \in H_0^1(\Omega)$ to the following equations:

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

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 [FMP15] 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 $\alpha_1 > \sigma_{\text{Pcr}}^2 (\|a_2\|_{L^\infty(\Omega)} + \|a_3\|_{L^\infty(\Omega)}) \geq 0$. Here, $\sigma_{\text{Pcr}} > 0$ denotes the constant in the Poincaré inequality $\|\cdot\|_{H^1(\Omega)} \leq \sigma_{\text{Pcr}} \|\cdot\|_{H^1(\Omega)}$ on $H_0^1(\Omega)$.

Definition 2.1. *We introduce the bilinear form:*

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

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

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

The assumptions on the PDE coefficients imply that the bilinear form $a(\cdot, \cdot)$ is continuous and coercive, cf. L.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 $\mathcal{T} \subseteq \text{Pow}(\Omega)$ 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 $F_T : \mathbb{R}^d \rightarrow \mathbb{R}^d$ 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., $\bigcup_{T \in \mathcal{T}} \bar{T} = \bar{\Omega}$. Finally, a mesh must be regular in the sense of [Cia78].*

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 D.2.8.

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

$$\mathcal{T}(B) := \{T \in \mathcal{T} \mid \bar{T} \cap \bar{B} \neq \emptyset\} \subseteq \mathcal{T}$$

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 $\sigma_{\text{shp}} \geq 1$ such that

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

Definition 2.3. *We define the mesh metric*

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

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

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

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

We refer to L.3.16 for some basic properties of the mesh metric.

Compared to [FMP15], we consider a more general class of meshes. Here, the crucial property is the so called *locally bounded cardinality* defined in the following D.2.4. Note that both *uniform* and *graded* meshes have this property, cf. Section 3.2.

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

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

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*

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

where $\mathbb{P}^p(\hat{T}) := \text{span}\{\hat{T} \ni x \mapsto x^q \mid \|q\|_1 \leq p\}$ 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 $p \geq 1$ and $N := \dim \mathbb{S}_0^{p,1}(\mathcal{T})$. A basis $\{\varphi_1, \dots, \varphi_N\} \subseteq \mathbb{S}_0^{p,1}(\mathcal{T})$ has a system of (local) dual functions $\{\lambda_1, \dots, \lambda_N\} \subseteq L^2(\Omega)$, if, for all $n, m \in \{1, \dots, N\}$ and $\mathbf{x} \in \mathbb{R}^N$, there hold the relations*

$$\langle \varphi_n, \lambda_m \rangle_{L^2(\Omega)} = \delta_{nm}, \quad \left\| \sum_{m=1}^N \mathbf{x}_m \lambda_m \right\|_{L^2(\Omega)} \lesssim h_{\min, \mathcal{T}}^{-d/2} \|\mathbf{x}\|_2.$$

The implied constant may only depend on d , p , and the shape regularity of the mesh \mathcal{T} .

Remark 2.7. *Note that we do not assume local basis functions φ_n , i.e., $\text{supp}_{\mathcal{T}}(\varphi_n) = \mathcal{T}$ is allowed. On the other hand, the dual functions λ_n should have local supports in order to guarantee competitive memory requirements for the \mathcal{H} -matrices (cf. R.2.12). Furthermore, the specific exponent of $h_{\min, \mathcal{T}}^{-d/2}$ in the stability bound is not crucial, as it only affects the exponent of the prefactor $N^{\sigma_{\text{card}}+2}$ in T.2.13.*

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 physical subdomains of Ω , which is captured in the following definition.

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

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

Recall from Section 2.2 that $\mathcal{T}(B) \subseteq \mathcal{T}$ is the patch of a physical subdomain $B \subseteq \mathbb{R}^d$ and that $\mathcal{T}(\mathcal{B}) \subseteq \mathcal{T}$ is the patch of a cluster $\mathcal{B} \subseteq \mathcal{T}$. Now, 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 $\mathcal{T} \subseteq \text{Pow}(\Omega)$ 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 Section 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

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

Again, existence and uniqueness of a solution $u \in \mathbb{S}_0^{p,1}(\mathcal{T})$ follow from L.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 D.2.1 and the basis functions $\varphi_n \in \mathbb{S}_0^{p,1}(\mathcal{T})$ from D.2.6 compose the governing system matrix.

Definition 2.9. *We define the system matrix*

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

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

2.5. Hierarchical matrices.

Definition 2.10. *A subset $\mathbb{P} \subseteq \text{Pow}(\{1, \dots, N\}) \times \text{Pow}(\{1, \dots, N\})$ is called a block partition, if*

$$\bigcup_{(I,J) \in \mathbb{P}} I \times J = \{1, \dots, N\} \times \{1, \dots, N\}.$$

Let $\sigma_{\text{adm}}, \sigma_{\text{small}} > 0$. A block partition \mathbb{P} is called admissible, if it can be split into parts

$$\mathbb{P} = \mathbb{P}_{\text{adm}} \dot{\cup} \mathbb{P}_{\text{small}}$$

with

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

Typically, an admissible block partition \mathbb{P} is constructed in two stages:

First, the indices $I_{\text{root}} := \{1, \dots, N\}$ are split up into a (*hierarchical*) *cluster tree* $\mathbb{T}_N := (\mathbb{T}_N^{(L)})_{L \geq 1}$. The first level is $\mathbb{T}_N^{(1)} := \{I_{\text{root}}\}$. Then, given the level $\mathbb{T}_N^{(L)}$, all $I \in \mathbb{T}_N^{(L)}$ with $\#I > \sigma_{\text{small}}$ are split in the form $I = I_1 \dot{\cup} I_2$ with $I_1 \neq \emptyset \neq I_2$ via a predefined *clustering strategy* $I \mapsto (I_1, I_2)$. (See, e.g., [Hac15] for some examples of such clustering strategies.) The combined set of all such children defines the next layer, $\mathbb{T}_N^{(L+1)}$. Clearly, this process stops after a finite number of layers denoted by $\text{depth}(\mathbb{T}_N)$.

Second, the matrix indices $I_{\text{root}} \times I_{\text{root}}$ are split up into a (*hierarchical*) *block cluster tree* $\mathbb{T}_{N \times N} := (\mathbb{T}_{N \times N}^{(L)})_{L \geq 1}$. Here, the first level is $\mathbb{T}_{N \times N}^{(1)} := \{(I_{\text{root}}, I_{\text{root}})\}$. Then, given the level $\mathbb{T}_{N \times N}^{(L)}$, all $(I, J) \in \mathbb{T}_{N \times N}^{(L)}$ with $\text{diam}_{\mathcal{T}}(\mathcal{T}(I)) > \sigma_{\text{adm}} \text{dist}_{\mathcal{T}}(\mathcal{T}(I), \mathcal{T}(J))$ are split into the children $(I_1, J_1), (I_1, J_2), (I_2, J_1), (I_2, J_2)$, where $I = I_1 \dot{\cup} I_2$ and $J = J_1 \dot{\cup} J_2$ as before. Again, all these children are collected in the layer $\mathbb{T}_{N \times N}^{(L+1)}$. Finally, the block partition \mathbb{P} is just the set of all leaves of $\mathbb{T}_{N \times N}$.

Definition 2.11. *Let \mathbb{P} be an admissible block partition and $r \in \mathbb{N}$ a given block rank bound. We define the set of \mathcal{H} -matrices by*

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

Remark 2.12. *By [Hac15, Lemma 6.13], the memory requirements to store an \mathcal{H} -matrix $\mathbf{B} \in \mathcal{H}(\mathbb{P}, r)$ can be bounded by the quantity $C_{\text{sparse}}(\mathbb{T}_{N \times N})(\sigma_{\text{small}} + r)\text{depth}(\mathbb{T}_N)N$, where $C_{\text{sparse}}(\mathbb{T}_{N \times N}) > 0$ denotes the so-called sparsity constant.*

In [GHLB04], the authors present a geometrically balanced clustering strategy that guarantees the upper bounds $C_{\text{sparse}}(\mathbb{T}_{N \times N}) \lesssim 1$ and $\text{depth}(\mathbb{T}_N) \lesssim \ln(h_{\text{min}, \mathcal{T}}^{-1})$. Using the relation $h_{\text{min}, \mathcal{T}} \gtrsim h_{\mathcal{T}}^{\sigma_{\text{card}}}$ from D.2.4 for meshes with locally bounded cardinality, we can conclude $\text{depth}(\mathbb{T}_N) \lesssim \ln(N)$. In particular, we get an overall bound of $\mathcal{O}(rN \ln N)$ for the memory requirements to store the matrix \mathbf{B} .

Note that this line of reasoning implicitly assumes that the dual functions $\lambda_n \in L^2(\Omega)$ from D.2.6 have local supports. More precisely, we need $\text{supp}_{\mathcal{T}}(\lambda_n) \subseteq \mathcal{T}(T_n)$ for some $T_n \in \mathcal{T}$ and have to ensure that these characteristic elements T_n do not coincide too frequently, i.e. $\#\{n \mid T_n = T\} \lesssim 1$ for all elements $T \in \mathcal{T}$.

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 by hierarchical matrices.

Theorem 2.13. *Let $\mathcal{T} \subseteq \text{Pow}(\Omega)$ be a mesh of locally bounded cardinality for some $\sigma_{\text{card}} \geq 1$ in the sense of D.2.4 and $\{\varphi_1, \dots, \varphi_N\} \subseteq \mathbb{S}_0^{p,1}(\mathcal{T})$ a basis that has a system of local dual functions (see D.2.6). Let $a(\cdot, \cdot)$ be the elliptic bilinear form from D.2.1 and $\mathbf{A} \in \mathbb{R}^{N \times N}$ be the corresponding Galerkin stiffness matrix (D.2.9). Finally, let \mathbb{P} be an admissible block partition as in D.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 $\mathbf{B} \in \mathcal{H}(\mathbb{P}, r)$ with*

$$\|\mathbf{A}^{-1} - \mathbf{B}\|_2 \lesssim N^{\sigma_{\text{card}}+2} \exp(-\sigma_{\text{exp}} r^{1/(d\sigma_{\text{card}}+1)}).$$

Under additional assumptions on the block partition \mathbb{P} , one can reduce the prefactor from $N^{\sigma_{\text{card}}+2}$ to $\ln(N)N^{\sigma_{\text{card}}}$, see R.3.13. As shown in Section 3.2, *uniform* and *algebraically graded* meshes have locally bounded cardinality. In particular, we immediately get the following corollary.

Corollary 2.14. *Let $\mathcal{T} \subseteq \text{Pow}(\Omega)$ be an algebraically graded mesh with grading exponent $\alpha \geq 1$ (see D.3.4). Then T.2.13 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 [FMP15] 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 D.2.4 and D.2.6. In Section 3.2, we present two types of meshes with locally bounded cardinality, namely *uniform* and *graded* meshes. The fact that

every uniform mesh has locally bounded cardinality will be used during our proof in T.3.32. The locally bounded cardinality of graded meshes shows that T.2.13 is applicable for graded meshes in the sense of D.3.4.

Then, in Section 3.3, we present a practical choice for the dual functions $\lambda_n \in L^2(\Omega)$ from D.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 T.2.13 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 \mathbf{A}^{-1} . Since \mathbf{A}^{-1} represents the act of solving the discretized model problem, it is only natural that the corresponding *discrete solution operator* $S_{\mathcal{T}} : L^2(\Omega) \rightarrow \mathbb{S}_0^{p,1}(\mathcal{T})$ will be involved. Additionally, this endeavor requires the dual functions $\lambda_n \in L^2(\Omega)$ mentioned earlier. We present the explicit formula for \mathbf{A}^{-1} at the end of Section 3.4.

3) In Section 3.5 we use this formula to go from the “matrix level” to the “function level”: Initially, we reduce the problem of approximating \mathbf{A}^{-1} as a whole to the problem of approximating $\mathbf{A}^{-1}|_{I \times J}$ for each admissible block $(I, J) \in \mathbb{P}_{\text{adm}}$. (The small blocks $\mathbb{P}_{\text{small}}$ are irrelevant in this 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 $\text{supp}_{\mathcal{T}}(f) \subseteq \mathcal{D}$? More precisely, we want to achieve the bounds (for some fixed $\kappa \geq 1$)

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

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 Section 3.6.

5) In Section 3.7 we investigate an important class of functions for our analysis, the spaces of *locally discrete harmonic functions* $\mathbb{S}_{\text{harm}}(\mathcal{B}) \subseteq \mathbb{S}_0^{p,1}(\mathcal{T})$. 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 Section 3.8 we construct the *single- and multi-step coarsening operators*. For any given $u \in \mathbb{S}_{\text{harm}}(\mathcal{B}^{\delta})$ on the inflated cluster $\mathcal{B}^{\delta} \supseteq \mathcal{B}$, the single-step coarsening operator $Q_{\mathcal{B}}^{\delta}$ produces a “coarse” approximation $Q_{\mathcal{B}}^{\delta}u \in \mathbb{S}_{\text{harm}}(\mathcal{B})$ 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 Section 3.9 we merely put all the pieces together and finish the proof of T.2.13.

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

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

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

Proof. Both estimates follow from the relation $\sum_{T \in \mathcal{B}} h_T^d \approx \sum_{T \in \mathcal{B}} |T| = |\bigcup \mathcal{B}|$ with appropriate $\mathcal{B} \subseteq \mathcal{T}$. \square

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

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

Using L.3.1 we immediately get the following result:

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

Definition 3.4 (Mesh graded towards Γ). *Let $\mathcal{T} \subseteq \text{Pow}(\Omega)$ be a mesh and $\Gamma \subseteq \mathbb{R}^d$ satisfy $\Gamma \subseteq \mathbb{R}^d \setminus T$ 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 graded towards Γ with parameters α, H , if there holds*

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

Here, x_T denotes the incenter of the element T and $\text{dist}_2(x_T, \Gamma) = \inf_{\gamma \in \Gamma} \|x_T - \gamma\|_2$ 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 Section 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 regain the optimal order of convergence. To a large extent, the shape of Γ is irrelevant for our analysis. We only require that the mesh resolve Γ , i.e., the mesh can only be graded towards points/lines that are part of the mesh skeleton.

Lemma 3.5. *Let $\mathcal{T} \subseteq \text{Pow}(\Omega)$ 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 $\sigma_{\text{card}} = \alpha$.*

Proof. We start with the bounds for $h_{\mathcal{T}}$ and $h_{\min, \mathcal{T}}$: For every $T \in \mathcal{T}$, we know from D.2.2 that $\text{Ball}_2(x_T, \sigma_{\text{shp}}^{-1} h_T) \subseteq T$. Combining this with the assumption $\Gamma \subseteq T^c$ from D.3.4 yields $\text{dist}_2(x_T, \Gamma) \geq h_T / \sigma_{\text{shp}}$. We conclude $h_T \approx \text{dist}_2(x_T, \Gamma)^{1-1/\alpha} H \gtrsim h_T^{1-1/\alpha} H$ and ultimately $h_{\min, \mathcal{T}} \gtrsim H^\alpha$. On the other hand, we have the bound $h_T \approx \text{dist}_2(x_T, \Gamma)^{1-1/\alpha} H \leq \sup_{x \in \Omega} \text{dist}_2(x, \Gamma)^{1-1/\alpha} H \lesssim H$ and thus $h_{\mathcal{T}} \lesssim H$.

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

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

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

In particular, both cases lead to the estimate

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

which concludes the proof. \square

3.3. Examples of 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 classic *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 dual system $\{\lambda_1, \dots, \lambda_N\} \subseteq L^2(\Omega)$ in the sense of D.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 $T_n \in \text{supp}_{\mathcal{T}}(\varphi_n) \subseteq \mathcal{T}(T_n)$.
- 2) *Simple structure:* There exists a basis of *shape functions* $\{\hat{\varphi}_1, \dots, \hat{\varphi}_L\} \subseteq \mathbb{P}^p(\hat{T})$, which determines the shape of the basis elements. More precisely, for every $n \in \{1, \dots, N\}$ and every $T \in \text{supp}_{\mathcal{T}}(\varphi_n)$, there exists an index $l(n, T) \in \{1, \dots, L\}$ such that $\varphi_n|_T = \hat{\varphi}_{l(n, T)} \circ F_T^{-1}$.
- 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 $T \in \text{supp}_{\mathcal{T}}(\varphi_n) \cap \text{supp}_{\mathcal{T}}(\varphi_m)$, there holds $l(n, T) \neq l(m, T)$.

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:

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

Now, 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. $\langle \hat{\varphi}_l, \hat{\lambda}_k \rangle_{L^2(\hat{T})} = \delta_{lk}$ for all $l, k \in \{1, \dots, L\}$. Then, the *dual function* $\lambda_n \in \mathbb{S}^{p,0}(\mathcal{T}) \subseteq L^2(\Omega)$ is defined in a piecewise manner: For every $T \neq T_n$, we set $\lambda_n|_T := 0$, whereas

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

Lemma 3.6. *For all $n, m \in \{1, \dots, N\}$ and $\mathbf{x} \in \mathbb{R}^N$, there holds*

$$\langle \varphi_n, \lambda_m \rangle_{L^2(\Omega)} = \delta_{nm}, \quad \left\| \sum_{m=1}^N \mathbf{x}_m \lambda_m \right\|_{L^2(\Omega)} \lesssim h_{\min, \mathcal{T}}^{-d/2} \|\mathbf{x}\|_2.$$

Proof. Let $n, m \in \{1, \dots, N\}$. If $T_m \notin \text{supp}_{\mathcal{T}}(\varphi_n)$, we have $m \neq n$ and therefore $\langle \varphi_n, \lambda_m \rangle_{L^2(\Omega)} = 0 = \delta_{nm}$. In the remaining case $T_m \in \text{supp}_{\mathcal{T}}(\varphi_n)$ we get

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

Next, recall that $|T| \approx h_T^d$ for every element T in a shape-regular mesh \mathcal{T} . For all $m \in \{1, \dots, N\}$, we compute

$$\|\lambda_m\|_{L^2(\Omega)} = |\det \nabla F_{T_m}|^{-1} \|\hat{\lambda}_{l(m, T_m)} \circ F_{T_m}^{-1}\|_{L^2(T_m)} = |\hat{T}|^{1/2} |T_m|^{-1/2} \|\hat{\lambda}_{l(m, T_m)}\|_{L^2(\hat{T})} \approx h_{T_m}^{-d/2}.$$

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. In particular, there must hold $\#ms(T) \lesssim 1$. Now, for every $\mathbf{x} \in \mathbb{R}^N$ and every $T \in \mathcal{T}$, we obtain

$$\left\| \sum_{m=1}^N \mathbf{x}_m \lambda_m \right\|_{L^2(T)}^2 = \left\| \sum_{m \in ms(T)} \mathbf{x}_m \lambda_m \right\|_{L^2(T)}^2 \leq \left(\sum_{m \in ms(T)} \|\lambda_m\|_{L^2(\Omega)}^2 \right) \left(\sum_{m \in ms(T)} \mathbf{x}_m^2 \right) \lesssim h_T^{-d} \sum_{m \in ms(T)} \mathbf{x}_m^2.$$

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

3.4. A representation formula for the inverse system matrix. In this subsection, we develop a representation formula for \mathbf{A}^{-1} in terms of three linear operators: Recall that \mathbf{A}^{-1} represents the action of solving the discrete model problem, so there must be a fundamental connection to the *discrete solution operator* $S_{\mathcal{T}} : L^2(\Omega) \rightarrow \mathbb{S}_0^{p,1}(\mathcal{T})$. Additionally, we need a way to turn coefficient vectors $\mathbf{f} \in \mathbb{R}^N$ 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 D.2.6 and the corresponding *coordinate mapping* $\Lambda : \mathbb{R}^N \rightarrow L^2(\Omega)$. Finally, the image $S_{\mathcal{T}} \Lambda \mathbf{f} \in \mathbb{S}_0^{p,1}(\mathcal{T})$ must be converted back to a vector in \mathbb{R}^N . A straightforward approach would be to use the inverse Φ^{-1} of the *coordinate mapping* $\Phi : \mathbb{R}^N \rightarrow \mathbb{S}_0^{p,1}(\mathcal{T})$ 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 $\Lambda^T : L^2(\Omega) \rightarrow \mathbb{R}^N$ instead.

First, let us recall the following classic result:

Lemma 3.7. *The bilinear form a from D.2.1 is coercive and continuous:*

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

The precise definitions of $S_{\mathcal{T}}$, Φ , and Λ are given in the following D.3.8.

Definition 3.8. *Let $a : H_0^1(\Omega) \times H_0^1(\Omega) \rightarrow \mathbb{R}$ the bilinear form from D.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*

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

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

Recall from Section 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 $\{\lambda_1, \dots, \lambda_N\} \subseteq L^2(\Omega)$ a dual system compliant with D.2.6. We denote the corresponding coordinate mappings by

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

We summarize the most important properties of Φ and Λ in the following lemma. As usual, we use the notation $\text{supp}(\mathbf{x}) := \{n \in \{1, \dots, N\} \mid \mathbf{x}_n \neq 0\}$ for the *support* of a vector $\mathbf{x} \in \mathbb{R}^N$. Furthermore, recall from D.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 Φ^{-1} . More precisely, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$ and all $v \in \mathbb{S}_0^{p,1}(\mathcal{T})$, there hold the duality/inversion formulae

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

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

$$\text{supp}_{\mathcal{T}}(\Lambda \mathbf{x}) \subseteq \mathcal{T}(\text{supp}(\mathbf{x})), \quad \|\Lambda^T v\|_{l^2(I)} \leq \|\Lambda\| \|v\|_{L^2(\mathcal{T}(I))}.$$

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

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

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

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

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

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

$$\forall \mathbf{x} \in \mathbb{R}^N : \quad \text{supp}_{\mathcal{T}}(\Lambda \mathbf{x}) = \text{supp}_{\mathcal{T}} \left(\sum_{n \in \text{supp}(\mathbf{x})} \mathbf{x}_n \lambda_n \right) \subseteq \bigcup_{n \in \text{supp}(\mathbf{x})} \text{supp}_{\mathcal{T}}(\lambda_n) \stackrel{D.2.8}{=} \mathcal{T}(\text{supp}(\mathbf{x})).$$

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,

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

which finishes the proof. \square

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

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

Proof. First, we establish a relationship between \mathbf{A} and a by means of the coordinate mapping Φ :

$$\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^N : \quad \langle \mathbf{A} \mathbf{x}, \mathbf{y} \rangle_2 \stackrel{D.2.9}{=} \sum_{n,m=1}^N a(\varphi_n, \varphi_m) \mathbf{x}_n \mathbf{y}_m \stackrel{D.3.9}{=} a(\Phi \mathbf{x}, \Phi \mathbf{y}).$$

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

$$\forall \mathbf{f}, \mathbf{y} \in \mathbb{R}^N : \quad \langle \mathbf{A} \Lambda^T S_{\mathcal{T}} \Lambda \mathbf{f}, \mathbf{y} \rangle_2 = a(\Phi \Lambda^T S_{\mathcal{T}} \Lambda \mathbf{f}, \Phi \mathbf{y}) = a(S_{\mathcal{T}} \Lambda \mathbf{f}, \Phi \mathbf{y}) \stackrel{D.3.8}{=} \langle \Lambda \mathbf{f}, \Phi \mathbf{y} \rangle_{L^2(\Omega)} = \langle \mathbf{f}, \mathbf{y} \rangle_2.$$

This readily implies the stated representation formula. \square

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 get rid of 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 $\mathbb{P} \subseteq \text{Pow}(\{1, \dots, N\}) \times \text{Pow}(\{1, \dots, N\})$ be a block partition. Then there holds the estimate*

$$\forall \mathbf{B} \in \mathbb{R}^{N \times N} : \quad \|\mathbf{B}\|_2 \leq N^2 \cdot \max_{(I,J) \in \mathbb{P}} \|\mathbf{B}|_{I \times J}\|_2.$$

Proof. The statement follows from

$$\forall \mathbf{x} \in \mathbb{R}^N : \quad \|\mathbf{B}\mathbf{x}\|_2^2 = \langle \mathbf{B}\mathbf{x}, \mathbf{B}\mathbf{x} \rangle_2 = \sum_{(I,J) \in \mathbb{P}} \langle \mathbf{B}|_{I \times J} \mathbf{x}|_J, (\mathbf{B}\mathbf{x})|_I \rangle_2 \leq \left(\max_{(I,J) \in \mathbb{P}} \|\mathbf{B}|_{I \times J}\|_2 \right) \#\mathbb{P} \|\mathbf{x}\|_2 \|\mathbf{B}\mathbf{x}\|_2$$

and the bound $\#\mathbb{P} \leq N^2$, which is valid for *any* partition \mathbb{P} of $\{1, \dots, N\} \times \{1, \dots, N\}$. \square

Remark 3.13. *The constant $\mathcal{O}(N^2)$ in the upper bound is far from optimal. If one assumes a block partition \mathbb{P} stemming from a hierarchical cluster tree \mathbb{T}_N , then it can be reduced to $\mathcal{O}(\ln N)$: In [Hac15, Lemma 6.5.8], the author showed the bound $\|\mathbf{B}\|_2 \leq C_{\text{sparse}}(\mathbb{T}_{N \times N}) \text{depth}(\mathbb{T}_N) \max_{(I,J) \in \mathbb{P}} \|\mathbf{B}|_{I \times J}\|_2$ with the sparsity constant $C_{\text{sparse}}(\mathbb{T}_{N \times N})$ and the depth of the cluster tree $\text{depth}(\mathbb{T}_N)$. Again, due to [GHLB04], one can achieve $C_{\text{sparse}}(\mathbb{T}_{N \times N}) \lesssim 1$ and $\text{depth}(\mathbb{T}_N) \lesssim \ln(h_{\min, \mathcal{T}}^{-1}) \lesssim \ln(N)$ with a geometrically balanced cluster tree on any mesh satisfying $h_{\min, \mathcal{T}} \gtrsim h_{\mathcal{T}}^{\sigma_{\text{card}}}$.*

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

Lemma 3.14. *Let $(I, J) \in \mathbb{P}_{\text{adm}}$ and $V \subseteq L^2(\Omega)$ be a finite-dimensional subspace. Then, there exist matrices $\mathbf{X} \in \mathbb{R}^{I \times r}$ and $\mathbf{Y} \in \mathbb{R}^{J \times r}$ of size $r \leq \dim V$, such that there holds the error bound*

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

Proof. We use the transposed coordinate mapping $\Lambda^T : L^2(\Omega) \rightarrow \mathbb{R}^N$ from L.3.10 to define $\mathbf{V} := (\Lambda^T V)|_I \subseteq \mathbb{R}^I$. Note that $r := \dim \mathbf{V} \leq \dim V$. Next, let the columns of the matrix $\mathbf{X} \in \mathbb{R}^{I \times r}$ be an $l^2(I)$ -orthonormal basis of \mathbf{V} . In particular, the product $\mathbf{X}\mathbf{X}^T \in \mathbb{R}^{I \times I}$ represents the $l^2(I)$ -orthogonal projection from \mathbb{R}^I onto \mathbf{V} . Finally, set $\mathbf{Y} := (\mathbf{A}^{-1}|_{I \times J})^T \mathbf{X} \in \mathbb{R}^{J \times r}$.

Now, for every $\mathbf{f} \in \mathbb{R}^N$ with $\text{supp}(\mathbf{f}) \subseteq J$, we get the bound

$$\begin{aligned} \|(\mathbf{A}^{-1}|_{I \times J} - \mathbf{X}\mathbf{Y}^T)\mathbf{f}|_J\|_{l^2(I)} &= \|(I - \mathbf{X}\mathbf{X}^T)(\mathbf{A}^{-1}\mathbf{f})|_I\|_{l^2(I)} = \inf_{\mathbf{v} \in \mathbf{V}} \|(\mathbf{A}^{-1}\mathbf{f})|_I - \mathbf{v}\|_{l^2(I)} \\ &\stackrel{L.3.11}{=} \inf_{v \in V} \|\Lambda^T(S_{\mathcal{T}}\Lambda\mathbf{f} - v)\|_{l^2(I)} \stackrel{L.3.10}{\leq} \|\Lambda\| \cdot \inf_{v \in V} \|S_{\mathcal{T}}\Lambda\mathbf{f} - v\|_{L^2(\mathcal{T}(I))}. \end{aligned}$$

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

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

Let $\mathcal{B}, \mathcal{D} \subseteq \mathcal{T}$ with $0 < \text{diam}_{\mathcal{T}}(\mathcal{B}) \leq \sigma_{\text{adm}} \text{dist}_{\mathcal{T}}(\mathcal{B}, \mathcal{D})$ 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 $\text{supp}_{\mathcal{T}}(f) \subseteq \mathcal{D}$?

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 $\text{dist}_{\mathcal{T}}(\cdot, \cdot)$ from D.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 3.15. For every cluster $\mathcal{B} \subseteq \mathcal{T}$ and every radius $\delta \geq 0$, we introduce the inflated cluster

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

We summarize the most important facts about the mesh metric and inflated clusters in the subsequent lemma. We omit the elementary proofs, as they follow directly from the respective definitions.

Lemma 3.16. The mesh metric $\text{dist}_{\mathcal{T}}(\cdot, \cdot)$ from D.2.3 defines a metric on \mathcal{T} . There holds the triangle type inequality

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

For every element $T \in \mathcal{T}$ and every neighbor $S \in \mathcal{T}(T)$, the distance is bounded by $\text{dist}_{\mathcal{T}}(T, S) \leq \sigma_{\text{shp}} h_T$. On the other hand, for every $S \in \mathcal{T} \setminus \{T\}$, we have the lower bound $\text{dist}_{\mathcal{T}}(T, S) \geq \sigma_{\text{shp}}^{-1}(h_T + h_S)$. Additionally, for every cluster $\mathcal{B} \subseteq \mathcal{T}$, there holds $h_{\mathcal{B}} \leq \max\{h_{\min, \mathcal{B}}, \sigma_{\text{shp}} \text{diam}_{\mathcal{T}}(\mathcal{B})\}$.

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

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

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 $\mathcal{T}(\mathcal{B}) \subseteq \mathcal{B}^{\sigma_{\text{shp}} h_{\mathcal{B}}}$. We conclude this summary with the bounds $\text{diam}_{\mathcal{T}}(\mathcal{B}^\delta) \leq \text{diam}_{\mathcal{T}}(\mathcal{B}) + 2\delta$ and $h_{\mathcal{B}^\delta} \leq \max\{h_{\mathcal{B}}, \sigma_{\text{shp}} \delta\}$.

For the construction of the cut-off function $\kappa_{\mathcal{B}}^\delta$ in L.3.19 we will use a variant of the classic *Clément operator*, [Clé75].

Definition 3.17. Let $\mathcal{N} \subseteq \bar{\Omega}$ be the nodes of the mesh \mathcal{T} and denote by $\{b_N \mid N \in \mathcal{N}\} \subseteq \mathbb{S}^{1,1}(\mathcal{T})$ the well-known hat-functions, i.e. $b_N(M) = \delta_{NM}$. We write $\langle v \rangle_T := |T|^{-1} \int_T v \, dx \in \mathbb{R}$ for the mean value of a function $v \in L^2(\Omega)$ on an element $T \in \mathcal{T}$. Now, the Clément operator $J_{\mathcal{T}} : L^2(\Omega) \rightarrow \mathbb{S}^{1,1}(\mathcal{T})$ 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 3.18. 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 $\text{supp}_{\mathcal{T}}(J_{\mathcal{T}} v) \subseteq \mathcal{T}(\text{supp}(v))$. 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

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

The discretized model problem $a(u, v) = \langle f, v \rangle_{L^2(\Omega)}$ was phrased in terms of *global* functions $u, v \in \mathbb{S}_0^{p,1}(\mathcal{T})$. But if we plug in a function v with local support, e.g., $\text{supp}_{\mathcal{T}}(v) \subseteq \mathcal{B}$ 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 3.19. 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

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

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

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

From the definition we immediately get $\text{supp}_{\mathcal{T}}(\kappa) \subseteq \mathcal{T}(\mathcal{B})^\varepsilon$ 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 L.3.16 to the clusters $\{T\}, \{S\}, \mathcal{T}(\mathcal{B})$ and

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

$$|\kappa|_T - \kappa|_S| \leq \frac{|\text{dist}_{\mathcal{T}}(T, \mathcal{T}(\mathcal{B})) - \text{dist}_{\mathcal{T}}(S, \mathcal{T}(\mathcal{B}))|}{\varepsilon} \leq \frac{\text{dist}_{\mathcal{T}}(T, S)}{\varepsilon} \stackrel{L.3.16}{\lesssim} \frac{h_T}{\varepsilon} \approx \frac{h_T}{\delta}.$$

We use the Clément operator $J_{\mathcal{T}} : L^2(\Omega) \rightarrow \mathbb{S}^{1,1}(\mathcal{T})$ from D.3.17 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

$$\text{supp}_{\mathcal{T}}(\kappa_{\mathcal{B}}^{\delta}) \stackrel{L.3.18}{\subseteq} \mathcal{T}(\text{supp}_{\mathcal{T}}(\kappa)) \subseteq \mathcal{T}(\mathcal{T}(\mathcal{B})^{\varepsilon}) \stackrel{L.3.16}{\subseteq} \mathcal{B}^{(1+\sigma_{\text{shp}}^2)(\sigma_{\text{shp}}h_{\mathcal{B}}+\varepsilon)} \subseteq \mathcal{B}^{2\sigma_{\text{shp}}^3h_{\mathcal{B}}+\delta/2} \stackrel{\delta \gtrsim h_{\mathcal{B}}}{\subseteq} \mathcal{B}^{\delta}.$$

From L.3.18 and $\kappa|_{\mathcal{T}(\mathcal{B})} \equiv 1$ we get $\kappa_{\mathcal{B}}^{\delta}|_{\mathcal{B}} \equiv 1$. Moreover, $0 \leq \kappa \leq 1$ yields $0 \leq \kappa_{\mathcal{B}}^{\delta} \leq 1$. This implies, in particular, $\|\kappa_{\mathcal{B}}^{\delta}\|_{L^{\infty}(\Omega)} \leq 1 \lesssim \delta^{-1}$, where we used the assumption $\delta \lesssim 1$. The remaining bound $|\kappa_{\mathcal{B}}^{\delta}|_{W^{1,\infty}(\Omega)} \lesssim \delta^{-1}$ follows from

$$\forall T \in \mathcal{T} : \quad h_T |\kappa_{\mathcal{B}}^{\delta}|_{W^{1,\infty}(T)} \stackrel{L.3.18}{\lesssim} \max_{S \in \mathcal{T}(T)} |\kappa|_T - \kappa|_S| \lesssim \frac{h_T}{\delta}.$$

This finishes the proof. \square

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 3.20. *Let $p \geq 1$ and denote by $\hat{I}^p : C^0(\hat{T}) \rightarrow \mathbb{P}^p(\hat{T})$ the (local) Lagrange interpolation operator on the reference element \hat{T} . The (global) Lagrange interpolation operator $I_{\mathcal{T}}^p : C_{\text{pw}}^0(\mathcal{T}) \rightarrow \mathbb{S}^{p,0}(\mathcal{T})$ is defined in a piecewise manner: For every $v \in C_{\text{pw}}^0(\mathcal{T})$ 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 inverse inequality (see, e.g., [DFG⁺01]).

Lemma 3.21. *Let $k \geq l \geq 0$, $q \in [1, \infty]$ and $p \geq 0$. 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*

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

The properties of the Lagrange interpolation operator $I_{\mathcal{T}}^p$ are very similar to those of the Clément operator $J_{\mathcal{T}}$ from D.3.17. For the sake of completeness, we include them in the following lemma.

Lemma 3.22. *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 $v \in C_{\text{pw}}^0(\mathcal{T})$ 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 $v \in C^0(\bar{\Omega})$, there holds $I_{\mathcal{T}}^p v \in \mathbb{S}^{p,1}(\mathcal{T})$. Similarly, if $v \in C^0(\bar{\Omega})$ 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 $\text{supp}_{\mathcal{T}}(I_{\mathcal{T}}^p v) \subseteq \text{supp}_{\mathcal{T}}(v)$. 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):*

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

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 L.3.21. \square

Remark 3.23. *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 D.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 3.24. *Let $\mathcal{B} \subseteq \mathcal{T}$ and $\delta > 0$ with $4\sigma_{\text{shp}}^3 h_{\mathcal{B}} \leq \delta \lesssim 1$ and denote by $\kappa_{\mathcal{B}}^{\delta} \in \mathbb{S}^{1,1}(\mathcal{T})$ the discrete cut-off function from L.3.19. Furthermore, denote by $I_{\mathcal{T}}^p : C_{\text{pw}}^0(\mathcal{T}) \rightarrow \mathbb{S}^{p,0}(\mathcal{T})$ the Lagrange interpolation operator from D.3.20. We define the discrete cut-off operator*

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

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

Lemma 3.25. *Let $\mathcal{B} \subseteq \mathcal{T}$ and $\delta > 0$ with $4\sigma_{\text{shp}}^3 h_{\mathcal{B}} \leq \delta \lesssim 1$. For all $v \in \mathbb{S}^{p,1}(\mathcal{T})$, the linear operator $K_{\mathcal{B}}^{\delta}$ has the cut-off property $\text{supp}_{\mathcal{T}}(K_{\mathcal{B}}^{\delta} v) \subseteq \mathcal{B}^{\delta}$ 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 L.3.22 and L.3.19. 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 L.3.22:

$$\sum_{l=0}^1 \delta^l \|K_{\mathcal{B}}^{\delta} v\|_{H^l(T)} \lesssim \sum_{l=0}^1 \delta^l \|\kappa_{\mathcal{B}}^{\delta} v\|_{H^l(T)} \lesssim \sum_{l=0}^1 \delta^l \sum_{i=0}^l |\kappa_{\mathcal{B}}^{\delta}|_{W^{l-i,\infty}(T)} \|v\|_{H^i(T)} \stackrel{\text{L.3.19}}{\lesssim} \sum_{l=0}^1 \delta^l \|v\|_{H^l(T)}.$$

□

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 Section 3.1, they are chosen for three main reasons: To begin with, they fit in seamlessly with the discrete solution operator $S_{\mathcal{T}} : L^2(\Omega) \rightarrow \mathbb{S}_0^{p,1}(\mathcal{T})$ from D.3.8. Furthermore, as specified in L.3.27, they are invariant with respect to the discrete cut-off operators $K_{\mathcal{B}}^{\delta} : \mathbb{S}^{p,1}(\mathcal{T}) \rightarrow \mathbb{S}^{p,1}(\mathcal{T})$ from D.3.24. 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*.

Definition 3.26. *For every $\mathcal{B} \subseteq \mathcal{T}$, we define the space of locally discrete harmonic functions*

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

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 $S_{\mathcal{T}} : L^2(\Omega) \rightarrow \mathbb{S}_0^{p,1}(\mathcal{T})$ and the discrete cut-off operators $K_{\mathcal{B}}^{\delta} : \mathbb{S}^{p,1}(\mathcal{T}) \rightarrow \mathbb{S}^{p,1}(\mathcal{T})$.

Lemma 3.27. *The spaces of locally discrete harmonic functions are nested in the sense*

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

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

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

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

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

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 $\text{supp}_{\mathcal{T}}(f) \subseteq \mathcal{D}$. Then, for every $v \in \mathbb{S}_0^{p,1}(\mathcal{T})$ with $\text{supp}_{\mathcal{T}}(v) \subseteq \mathcal{B}$, we have

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

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 $\text{supp}_{\mathcal{T}}(v) \subseteq \mathcal{B}$. Then,

$$\begin{aligned} a(K_{\mathcal{B}}^{\delta}u, v) &\stackrel{D.2.1}{=} \langle a_1 \nabla K_{\mathcal{B}}^{\delta}u, \nabla v \rangle_{L^2(\mathcal{B})} + \langle a_2 \cdot \nabla K_{\mathcal{B}}^{\delta}u, v \rangle_{L^2(\mathcal{B})} + \langle a_3 K_{\mathcal{B}}^{\delta}u, v \rangle_{L^2(\mathcal{B})} \\ &\stackrel{L.3.25}{=} \langle a_1 \nabla u, \nabla v \rangle_{L^2(\mathcal{B})} + \langle a_2 \cdot \nabla u, v \rangle_{L^2(\mathcal{B})} + \langle a_3 u, v \rangle_{L^2(\mathcal{B})} \\ &= a(u, v) \\ &= 0. \end{aligned}$$

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

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}^{δ} . Obviously, this can be true only 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 D.3.26. We can interpret the discrete Caccioppoli inequality as an improved version of the inverse inequality from L.3.21, 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 3.28. *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, an induction on $p \geq 1$ yields the following estimate: For every $\kappa \in \mathbb{S}^{1,0}(\mathcal{T})$, $u \in \mathbb{S}^{p,0}(\mathcal{T})$ and $T \in \mathcal{T}$,

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

In the base case $p = 1$, the second-order derivatives in $|\kappa^2 u|_{H^2(T)}$ can be computed explicitly. Since $\kappa, u \in \mathbb{P}^1(T)$, the terms containing $D^{\alpha} \kappa$ or $D^{\alpha} u$ 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 L.3.21 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 amenable to the induction hypothesis: $|\kappa^2(\partial_i u)|_{H^{p+1}(T)} \lesssim h_T^{1-p} |\kappa|_{W^{1,\infty}(\Omega)} (\|(\partial_i u) \nabla \kappa\|_{L^2(T)} + \|\kappa \nabla(\partial_i u)\|_{L^2(T)})$. These terms can be treated with the fact $\nabla \kappa \equiv \text{const}$, the identity $\kappa \nabla(\partial_i u) = \partial_i \kappa \nabla u - (\partial_i \kappa) \nabla u$ and the inverse inequality L.3.21 once again.

Now, let us turn our attention to the discrete Caccioppoli inequality itself. For this purpose, let $\mathcal{B} \subseteq \mathcal{T}$ and $\delta > 0$ with $4\sigma_{\text{shp}}^3 h_{\mathcal{B}} \leq \delta \lesssim 1$. We denote by $\kappa := \kappa_{\mathcal{B}}^{\delta} \in \mathbb{S}^{1,1}(\mathcal{T})$ the discrete cut-off function from L.3.19 and by $I_{\mathcal{T}}^p : C_{\text{pw}}^0(\mathcal{T}) \rightarrow \mathbb{S}^{p,0}(\mathcal{T})$ the Lagrange interpolation operator from D.3.20. Furthermore, let $u \in \mathbb{S}_{\text{harm}}(\mathcal{B}^{\delta})$. 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 $\text{supp}_{\mathcal{T}}(v) \subseteq \mathcal{B}^{\delta}$. From L.3.22 and L.3.19 we know that $v := I_{\mathcal{T}}^p(\kappa^2 u)$ satisfies both $v \in \mathbb{S}_0^{p,1}(\mathcal{T})$ and $\text{supp}_{\mathcal{T}}(v) \subseteq \text{supp}_{\mathcal{T}}(\kappa) \subseteq \mathcal{B}^{\delta}$, i.e., we can use v as said test function. This results in the following bound:

$$\begin{aligned} a(u, \kappa^2 u) &= a(u, (\text{id} - I_{\mathcal{T}}^p)(\kappa^2 u)) \\ &\stackrel{D.2.1}{\lesssim} \sum_{T \in \mathcal{B}^{\delta}} \|u\|_{H^1(T)} \|(\text{id} - I_{\mathcal{T}}^p)(\kappa^2 u)\|_{H^1(T)} \\ &\stackrel{L.3.22}{\lesssim} \sum_{T \in \mathcal{B}^{\delta}} \|u\|_{H^1(T)} h_T^p |\kappa^2 u|_{H^{p+1}(T)} \\ &\lesssim |\kappa|_{W^{1,\infty}(\Omega)} \sum_{T \in \mathcal{B}^{\delta}} h_T \|u\|_{H^1(T)} (\|u \nabla \kappa\|_{L^2(T)} + \|\kappa \nabla u\|_{L^2(T)}) \\ &\stackrel{L.3.21}{\lesssim} |\kappa|_{W^{1,\infty}(\Omega)} \|u\|_{L^2(\mathcal{B}^{\delta})} (\|u \nabla \kappa\|_{L^2(\Omega)} + \|\kappa \nabla u\|_{L^2(\Omega)}). \end{aligned}$$

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

$$\begin{aligned}
\|\kappa \nabla u\|_{L^2(\Omega)}^2 &\lesssim \langle a_1 \kappa \nabla u, \kappa \nabla u \rangle_{L^2(\Omega)} \\
&\stackrel{D.2.1}{=} a(u, \kappa^2 u) - 2\langle a_1 \kappa \nabla u, u \nabla \kappa \rangle_{L^2(\Omega)} - \langle a_2 \cdot \nabla u, \kappa^2 u \rangle_{L^2(\Omega)} - \langle a_3 u, \kappa^2 u \rangle_{L^2(\Omega)} \\
&\lesssim |\kappa|_{W^{1,\infty}(\Omega)} \|u\|_{L^2(\mathcal{B}^\delta)} (\|u \nabla \kappa\|_{L^2(\Omega)} + \|\kappa \nabla u\|_{L^2(\Omega)}) \\
&\quad + \|\kappa \nabla u\|_{L^2(\Omega)} \|u \nabla \kappa\|_{L^2(\Omega)} + \|\kappa \nabla u\|_{L^2(\Omega)} \|\kappa u\|_{L^2(\Omega)} + \|\kappa u\|_{L^2(\Omega)}^2 \\
&\stackrel{\forall \varepsilon > 0}{\leq} C_\varepsilon \|\kappa\|_{W^{1,\infty}(\Omega)}^2 \|u\|_{L^2(\mathcal{B}^\delta)}^2 + \varepsilon \|\kappa \nabla u\|_{L^2(\Omega)}^2.
\end{aligned}$$

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

$$|u|_{H^1(\mathcal{B})} \stackrel{\kappa|_{\mathcal{B}} \equiv 1}{\leq} \|\kappa \nabla u\|_{L^2(\Omega)} \lesssim \|\kappa\|_{W^{1,\infty}(\Omega)} \|u\|_{L^2(\mathcal{B}^\delta)} \stackrel{L.3.19}{\lesssim} \frac{1}{\delta} \|u\|_{L^2(\mathcal{B}^\delta)}.$$

This concludes the proof of the discrete Caccioppoli inequality. \square

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 Section 3.1. We design the so called *single-* and *multi-step coarsening operators*. For 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 T.3.32 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 $\Pi_{\mathcal{S}}^p : L^2(\Omega) \rightarrow \mathbb{S}^{p,0}(\mathcal{S})$ 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, L.3.28. 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 D.3.24. Finally, the combined operator $\Pi_{\mathcal{S}}^p K_{\mathcal{B}}^\delta : \mathbb{S}_{\text{harm}}(\mathcal{B}^\delta) \rightarrow \mathbb{S}^{p,0}(\mathcal{S})$ only lacks one more thing: It does not necessarily map into the space $\mathbb{S}_{\text{harm}}(\mathcal{B})$, which is a critical requirement, because we want to iterate the argument by plugging the remainder $\tilde{u} := u - Q_{\mathcal{B}}^\delta u$ of one single-step coarsening operator into another one. Thankfully, we can simply append the orthogonal projection $P_{\mathcal{B}} : L^2(\Omega) \rightarrow \mathbb{S}_{\text{harm}}(\mathcal{B})$ without losing any of the aforementioned properties.

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

Lemma 3.29. *Let $\mathcal{S}_0 \subseteq \text{Pow}(\Omega)$ be an arbitrary mesh and $(\mathcal{S}_l)_{l \in \mathbb{N}_0}$ be the corresponding sequence of uniform refinements. For every $H > 0$, there exists an $\mathcal{S} \in (\mathcal{S}_l)_{l \in \mathbb{N}_0}$ with $\sigma_{\text{shp}}(\mathcal{S}) = C(\Omega)$ and $C(\Omega)H \leq h_{\min, \mathcal{S}} \leq h_{\mathcal{S}} \leq H$. In particular, \mathcal{S} is uniform in the sense of D.3.2.*

Proof. There hold the relations $h_{\mathcal{S}_l} = 2^{-l} h_{\mathcal{S}_0}$ and $h_{\min, \mathcal{S}_l} = 2^{-l} h_{\min, \mathcal{S}_0}$. 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 $H < h_{\mathcal{S}_{L-1}} = 2^{-(L-1)} h_{\mathcal{S}_0} = 2 h_{\mathcal{S}_0} h_{\min, \mathcal{S}_0}^{-1} h_{\min, \mathcal{S}_L} = C(\Omega) h_{\min, \mathcal{S}_L}$. \square

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

Definition 3.30. Let $\mathcal{S} \subseteq \text{Pow}(\Omega)$ be a mesh, $p \geq 0$ and $\hat{\Pi}^p : L^2(\hat{T}) \rightarrow \mathbb{P}^p(\hat{T})$ the orthogonal projection on the reference element \hat{T} . The piecewise orthogonal projection $\Pi_{\mathcal{S}}^p : L^2(\Omega) \rightarrow \mathbb{S}^{p,0}(\mathcal{S})$ is defined in a piecewise manner: For every $v \in L^2(\Omega)$ and every $S \in \mathcal{S}$ we set

$$(\Pi_{\mathcal{S}}^p v)|_S := \hat{\Pi}^p(v \circ F_S) \circ F_S^{-1}.$$

In fact, $\Pi_{\mathcal{S}}^p$ coincides with the (global) orthogonal projection from $L^2(\Omega)$ onto the closed subspace $\mathbb{S}^{p,0}(\mathcal{S})$. The piecewise approach, however, results in desirable *local* properties and bounds.

Lemma 3.31. The linear operator $\Pi_{\mathcal{S}}^p$ 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 $(\Pi_{\mathcal{S}}^p v)|_{\mathcal{B}} = v|_{\mathcal{B}}$. Furthermore, $\Pi_{\mathcal{S}}^p$ preserves supports: For every $v \in L^2(\Omega)$, we have $\text{supp}_{\mathcal{S}}(\Pi_{\mathcal{S}}^p v) \subseteq \text{supp}_{\mathcal{S}}(v)$. Finally, for every $k \in \{0, \dots, p+1\}$, there hold the stability and error estimates

$$\forall v \in H_{\text{pw}}^k(\mathcal{S}) : \forall S \in \mathcal{S} : \quad \begin{aligned} \sum_{l=0}^k h_S^l |\Pi_{\mathcal{S}}^p v|_{H^l(S)} &\lesssim \sum_{l=0}^k h_S^l |v|_{H^l(S)}, \\ \sum_{l=0}^k h_S^l |(\text{id} - \Pi_{\mathcal{S}}^p)(v)|_{H^l(S)} &\lesssim h_S^k |v|_{H^k(S)}. \end{aligned}$$

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

Theorem 3.32. Let $\mathcal{T} \subseteq \text{Pow}(\Omega)$ 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

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

of rank

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

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

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

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 $K_{\mathcal{B}}^{\varepsilon} : \mathbb{S}^{p,1}(\mathcal{T}) \rightarrow \mathbb{S}^{p,1}(\mathcal{T})$ from D.3.24 with some carefully chosen parameter $\varepsilon > 0$. Second, we apply the piecewise orthogonal projection $\Pi_{\mathcal{S}}^p : L^2(\Omega) \rightarrow \mathbb{S}^{p,0}(\mathcal{S})$ from D.3.30 on some suitable mesh $\mathcal{S} \subseteq \text{Pow}(\Omega)$. Third, the result is mapped back into the space $\mathbb{S}_{\text{harm}}(\mathcal{B})$ via the orthogonal projection $P_{\mathcal{B}} : L^2(\Omega) \rightarrow \mathbb{S}_{\text{harm}}(\mathcal{B})$.

For the precise choice of ε and \mathcal{S} we have to distinguish between two cases: In the more involved case $\delta \geq 20\sigma_{\text{shp}}^7 h_{\mathcal{B}}$ we choose $\varepsilon := \delta/(5\sigma_{\text{shp}}^4) \geq 4\sigma_{\text{shp}}^3 h_{\mathcal{B}}$ and use the uniform mesh $\mathcal{S} \subseteq \text{Pow}(\Omega)$ from L.3.29 with $h_{\mathcal{S}} \approx h_{\min, \mathcal{S}} \approx H$, where the parameter $H > 0$ will be specified during the proof. In the degenerate case $\delta < 20\sigma_{\text{shp}}^7 h_{\mathcal{B}}$ we set $\varepsilon := 4\sigma_{\text{shp}}^3 h_{\mathcal{B}}$ and use the mesh $\mathcal{S} := \mathcal{T}$ itself.

We define the asserted operator as

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

The case $\delta \geq 20\sigma_{\text{shp}}^7 h_{\mathcal{B}}$: Let $u \in \mathbb{S}_{\text{harm}}(\mathcal{B}^{\delta})$. From L.3.16 we know that the parameter $\alpha := 4\sigma_{\text{shp}}^4 \varepsilon$ satisfies $4\sigma_{\text{shp}}^3 h_{\mathcal{B}^{\varepsilon}} \leq \alpha \lesssim 1$. In particular, we can apply the discrete Caccioppoli inequality to the set $\mathcal{B}^{\varepsilon}$ and the parameter α . Since $\delta \approx \alpha$, this gives the stability estimate for the cut-off operator $K_{\mathcal{B}}^{\varepsilon}$

$$\sum_{l=0}^1 \delta^l |K_{\mathcal{B}}^{\varepsilon} u|_{H^l(\Omega)} \stackrel{\text{L.3.25}}{\lesssim} \sum_{l=0}^1 \alpha^l |u|_{H^l(\mathcal{B}^{\varepsilon})} \stackrel{\text{L.3.28}}{\lesssim} \|u\|_{L^2(\mathcal{B}^{\varepsilon+\alpha})} \stackrel{\varepsilon+\alpha \leq \delta}{\leq} \|u\|_{L^2(\mathcal{B}^{\delta})}.$$

From L.3.27 and L.3.25 we know that $K_{\mathcal{B}}^{\varepsilon} u \in \mathbb{S}_{\text{harm}}(\mathcal{B})$, 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

$$\begin{aligned} \|u - Q_{\mathcal{B}}^{\delta} u\|_{L^2(\mathcal{B})} &= \|P_{\mathcal{B}} K_{\mathcal{B}}^{\varepsilon} u - P_{\mathcal{B}} \Pi_{\mathcal{S}}^p K_{\mathcal{B}}^{\varepsilon} u\|_{L^2(\mathcal{B})} \stackrel{\text{L.3.31}}{\leq} \|P_{\mathcal{B}}(\text{id} - \Pi_{\mathcal{S}}^p)(K_{\mathcal{B}}^{\varepsilon} u)\|_{L^2(\Omega)} \\ &\leq \|(\text{id} - \Pi_{\mathcal{S}}^p)(K_{\mathcal{B}}^{\varepsilon} u)\|_{L^2(\Omega)} \lesssim H |K_{\mathcal{B}}^{\varepsilon} u|_{H^1(\Omega)} \\ &\lesssim \frac{H}{\delta} \|u\|_{L^2(\mathcal{B}^{\delta})}. \end{aligned}$$

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

The case $\delta < 20\sigma_{\text{shp}}^7 h_{\mathcal{B}}$: Again let $u \in \mathbb{S}_{\text{harm}}(\mathcal{B}^\delta)$. Exploiting $\mathcal{S} = \mathcal{T}$ and L.3.25, 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 $\delta \geq 20\sigma_{\text{shp}}^7 h_{\mathcal{B}}$ the mesh \mathcal{S} is uniform and thus of locally bounded cardinality (cf. L.3.3). In the case $\delta < 20\sigma_{\text{shp}}^7 h_{\mathcal{B}}$ 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 $\delta \geq 20\sigma_{\text{shp}}^7 h_{\mathcal{B}}$ we have $h_{\mathcal{B}} + \varepsilon + \delta \lesssim \delta \approx H \approx h_{\min, \mathcal{S}} \leq h_{\mathcal{S}(B)}$ and in the case $\delta < 20\sigma_{\text{shp}}^7 h_{\mathcal{B}}$ we get $h_{\mathcal{B}} + \varepsilon + \delta \lesssim h_{\mathcal{B}} \leq h_{\mathcal{T}(B^\varepsilon)} = h_{\mathcal{S}(B)}$ as well.

Now, for every $u \in \mathbb{S}_{\text{harm}}(\mathcal{B}^\delta)$ we know from L.3.31 and L.3.25 that $\text{supp}_{\mathcal{S}}(\Pi_{\mathcal{S}}^p K_{\mathcal{B}}^\varepsilon u) \subseteq \text{supp}_{\mathcal{S}}(K_{\mathcal{B}}^\varepsilon u) \subseteq \mathcal{S}(B)$. This results in the estimate

$$\begin{aligned} \text{rank}(Q_{\mathcal{B}}^\delta) &\leq \dim \{v \in \mathbb{S}^{p,0}(\mathcal{S}) \mid \text{supp}_{\mathcal{S}}(v) \subseteq \mathcal{S}(B)\} \stackrel{\text{L.3.16}}{\approx} \#\mathcal{S}(B) \\ &\stackrel{\text{D.2.4}}{\lesssim} (1 + h_{\mathcal{S}(B)}^{-1} \text{diam}_{\mathcal{S}}(\mathcal{S}(B)))^{d\sigma_{\text{card}}} \stackrel{\text{L.3.16}}{\lesssim} (1 + h_{\mathcal{S}(B)}^{-1} (\text{diam}_{\mathcal{T}}(\mathcal{B}) + h_{\mathcal{B}} + \varepsilon))^{d\sigma_{\text{card}}} \\ &\stackrel{h_{\mathcal{B}} + \varepsilon + \delta \lesssim h_{\mathcal{S}(B)}}{\lesssim} (1 + \delta^{-1} \text{diam}_{\mathcal{T}}(\mathcal{B}))^{d\sigma_{\text{card}}}, \end{aligned}$$

which finishes the proof. \square

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

Theorem 3.33. *Let $\mathcal{T} \subseteq \text{Pow}(\Omega)$ 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*

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

of rank

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

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

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

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 $\mathcal{B} \subseteq \mathcal{B}_0 \subseteq \dots \subseteq \mathcal{B}_L \subseteq \mathcal{B}^{\delta L}$ inductively by $\mathcal{B}_0 := \mathcal{B}$ and $\mathcal{B}_{l+1} := (\mathcal{B}_l)^\delta$. Using the corresponding single-step coarsening operators $Q_l := Q_{\mathcal{B}_l}^\delta : \mathbb{S}_{\text{harm}}(\mathcal{B}_{l+1}) \longrightarrow \mathbb{S}_{\text{harm}}(\mathcal{B}_l)$ from T.3.32, we make the following definition:

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

Using the alternative representation $Q_{\mathcal{B}}^{\delta, L} u = -\sum_{\pi \in \{0,1\}^L \setminus \{0\}} (-Q_0)^{(\pi_0)} \circ \dots \circ (-Q_{L-1})^{(\pi_{L-1})}(u)$, we infer

$$\begin{aligned} \text{rank}(Q_{\mathcal{B}}^{\delta, L}) &\leq \sum_{l=0}^{L-1} \text{rank}(Q_l) \stackrel{\text{T.3.32}}{\lesssim} \sum_{l=0}^{L-1} (1 + \delta^{-1} \text{diam}_{\mathcal{T}}(\mathcal{B}_l))^{d\sigma_{\text{card}}} \\ &\stackrel{\text{L.3.16}}{\lesssim} \sum_{l=0}^{L-1} (1 + \delta^{-1} \text{diam}_{\mathcal{T}}(\mathcal{B}) + l)^{d\sigma_{\text{card}}} \leq (L + \delta^{-1} \text{diam}_{\mathcal{T}}(\mathcal{B}))^{d\sigma_{\text{card}} + 1}. \end{aligned}$$

Finally, the definition of $Q_{\mathcal{B}}^{\delta, L}$ was such that the error bound becomes elementary: For every $u \in \mathbb{S}_{\text{harm}}(\mathcal{B}^{\delta L})$, iteration of T.3.32 gives

$$\|u - Q_{\mathcal{B}}^{\delta, L} u\|_{L^2(\mathcal{B})} = \|(\text{id} - Q_0) \circ \dots \circ (\text{id} - Q_{L-1})(u)\|_{L^2(\mathcal{B}_0)} \leq 2^{-L} \|u\|_{L^2(\mathcal{B}^{\delta L})}.$$

\square

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 Section 3.1. After that, the proof of T.2.13 is just a matter of putting everything together.

Theorem 3.34. *Let $\mathcal{T} \subseteq \text{Pow}(\Omega)$ be a mesh of locally bounded cardinality and $\mathcal{B}, \mathcal{D} \subseteq \mathcal{T}$ clusters satisfying*

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

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

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

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

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

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

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

Here, $Q_{\mathcal{B}}^{\delta,L} : \mathbb{S}_{\text{harm}}(\mathcal{B}^{\delta L}) \rightarrow \mathbb{S}_{\text{harm}}(\mathcal{B})$ is the multi-step coarsening operator from T.3.33.

Using T.3.33 and the definition of δ , we can bound the dimension by

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

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

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

Then, L.3.27 implies $S_{\mathcal{T}}f \in \mathbb{S}_{\text{harm}}(\mathcal{B}^{\delta L})$ and ultimately

$$\inf_{v \in V_{\mathcal{B},\mathcal{D},L}} \|S_{\mathcal{T}}f - v\|_{L^2(\mathcal{B})} \leq \|S_{\mathcal{T}}f - Q_{\mathcal{B}}^{\delta,L}(S_{\mathcal{T}}f)\|_{L^2(\mathcal{B})} \stackrel{\text{T.3.33}}{\leq} 2^{-L} \|S_{\mathcal{T}}f\|_{L^2(\mathcal{B}^{\delta L})} \stackrel{\text{D.3.8}}{\lesssim} 2^{-L} \|f\|_{L^2(\mathcal{D})}.$$

□

We close this section with the proof of T.2.13.

Proof. Let $\mathbf{A} \in \mathbb{R}^{N \times N}$ be the matrix from D.2.9 and $r \in \mathbb{N}$ a given block rank bound. We define the asserted \mathcal{H} -matrix approximant $\mathbf{B} \in \mathbb{R}^{N \times N}$ to \mathbf{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 D.2.10 we know that $0 < \text{diam}_{\mathcal{T}}(\mathcal{B}) \leq \sigma_{\text{adm}} \text{dist}_{\mathcal{T}}(\mathcal{B}, \mathcal{D})$. Furthermore, let $C > 0$ be the constant from the dimension bound in T.3.34. 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, T.3.34 provides a subspace $V_{\mathcal{B},\mathcal{D},L} \subseteq \mathbb{S}_0^{p,1}(\mathcal{T}) \subseteq L^2(\Omega)$. We apply L.3.14 to the subspace $V_{\mathcal{B},\mathcal{D},L} \subseteq L^2(\Omega)$ and get matrices $\mathbf{X}_{I,J}^r \in \mathbb{R}^{I \times \tilde{r}}$ and $\mathbf{Y}_{I,J}^r \in \mathbb{R}^{J \times \tilde{r}}$ of size $\tilde{r} \leq \dim V_{\mathcal{B},\mathcal{D},L}$. We set

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

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

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

By D.2.11, we have $\mathbf{B} \in \mathcal{H}(\mathbb{P}, \tilde{r})$ with a block rank bound

$$\tilde{r} \leq \dim V_{\mathcal{B},\mathcal{D},L} \stackrel{\text{T.3.34}}{\leq} CL^{d\sigma_{\text{card}}+1} \leq r.$$

For the error we get

$$\begin{aligned}
\|\mathbf{A}^{-1} - \mathbf{B}\|_2 &\stackrel{L.3.12}{\leq} N^2 \cdot \max_{(I,J) \in \mathbb{P}_{\text{adm}}} \|\mathbf{A}^{-1}|_{I \times J} - \mathbf{X}_{I,J}^r (\mathbf{Y}_{I,J}^r)^T\|_2 \\
&\stackrel{L.3.14}{\leq} N^2 \|\Lambda\|^2 \cdot \max_{\substack{\mathcal{B}, \mathcal{D} \subseteq \mathcal{T} \\ \text{admissible}}} \sup_{\substack{f \in L^2(\Omega): \\ \text{supp}_{\mathcal{T}}(f) \subseteq \mathcal{D}}} \|f\|_{L^2(\mathcal{D})}^{-1} \cdot \inf_{v \in V_{\mathcal{B}, \mathcal{D}, L}} \|\mathcal{S}_{\mathcal{T}} f - v\|_{L^2(\mathcal{B})} \\
&\stackrel{T.3.34}{\lesssim} N^2 \|\Lambda\|^2 2^{-L} \\
&\lesssim N^2 \|\Lambda\|^2 \exp(-\sigma_{\text{exp}} r^{1/(d\sigma_{\text{card}}+1)}).
\end{aligned}$$

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

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

This concludes the proof of the main result, T.2.13. \square

4. NUMERICAL RESULTS

In this subsection, we illustrate the validity of T.2.13 by means of a numerical example:

For the geometry we choose the *L-shaped domain* $\Omega := ((0,1) \times (0,1)) \setminus ([1/2, 1] \times [1/2, 1]) \subseteq \mathbb{R}^2$ in two space dimensions. The PDE coefficients for the model problem from Section 2.1 are given by $a_1(x) = \begin{pmatrix} 10 & -1 \\ -1 & 1 \end{pmatrix}$, $a_2(x) := \begin{pmatrix} 10x_2 \\ 0 \end{pmatrix}$ and $a_3(x) := 1$. The mesh \mathcal{T} is *graded* in the sense of D.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 [GHLB04]. We choose the parameters $\sigma_{\text{adm}} := 2$ and $\sigma_{\text{small}} := 25$ (cf. D.2.10). For the rank bound we choose the range $r \in \{1, \dots, 50\}$.

Unfortunately, the \mathcal{H} -matrix approximant $\mathbf{B} \in \mathbb{R}^{N \times N}$ 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 values decomposition*: First, we compute the exact inverse $\mathbf{A}^{-1} \in \mathbb{R}^{N \times N}$ explicitly. Then, for every admissible block $(I, J) \in \mathbb{P}_{\text{adm}}$, we perform the singular values decomposition $\mathbf{A}^{-1}|_{I \times J} = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^T \in \mathbb{R}^{I \times J}$. Here, $\mathbf{U} \in \mathbb{R}^{I \times I}$, $\mathbf{V} \in \mathbb{R}^{J \times J}$ are orthogonal and $\mathbf{\Sigma} = \text{diag}(\sigma_1, \dots, \sigma_{\min\{\#I, \#J\}}) \in \mathbb{R}^{I \times J}$ contains the corresponding singular values $\sigma_1 \geq \dots \geq \sigma_{\min\{\#I, \#J\}} \geq 0$. Now, for the approximant we use $\mathbf{B}|_{I \times J} := \mathbf{U}_r \mathbf{\Sigma}_r \mathbf{V}_r^T \in \mathbb{R}^{I \times J}$, where $\mathbf{U}_r \in \mathbb{R}^{I \times r}$, $\mathbf{\Sigma}_r \in \mathbb{R}^{r \times r}$ and $\mathbf{V}_r \in \mathbb{R}^{J \times r}$ are the first r columns of \mathbf{U} , $\mathbf{\Sigma}$ and \mathbf{V} , respectively. Recall from the theory of singular values decompositions (e.g., [Hac15]) that

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

In particular, we end up with the following *computable* error bound (cf. [Hac15, Lemma 6.5.8])

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

The numerical example is implemented in MATLAB. For the inversion of the full matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$ we use MATLAB's built-in procedure `inv(...)`. For the singular values 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.

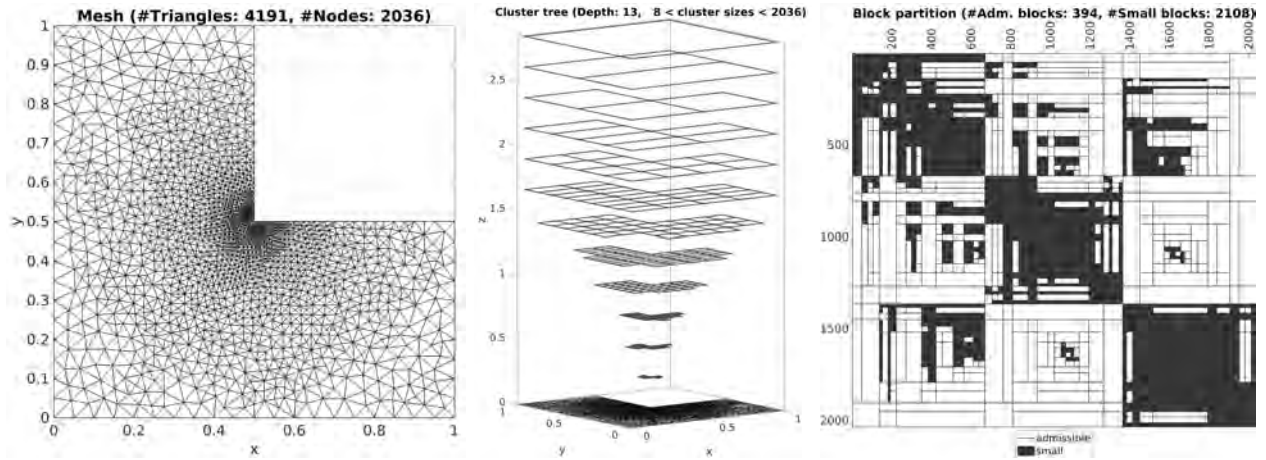


FIGURE 1. The mesh \mathcal{T} , the cluster tree \mathbb{T}_N and the block partition \mathbb{P} for $N \approx 2.000$ degrees of freedom.

In Figure 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 $N \rightarrow \infty$.

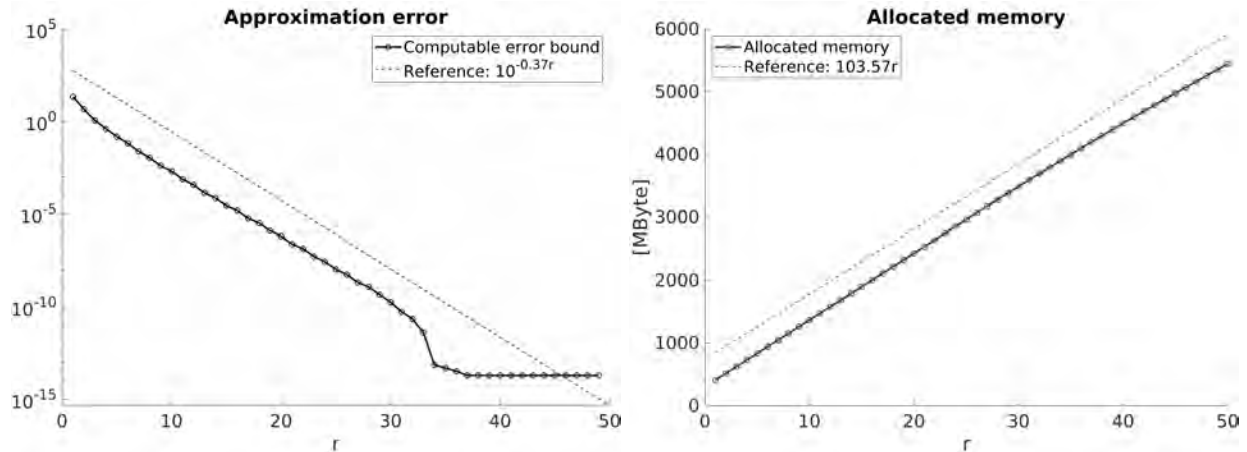


FIGURE 2. Approximation error and memory allocation for $N \approx 72.000$ degrees of freedom.

In Figure 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 T.2.13. 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.

REFERENCES

- [Beb05] M. Bebendorf, *Efficient inversion of Galerkin matrices of general second-order elliptic differential operators with nonsmooth coefficients*, Math. Comp. **74** (2005), 1179–1199.
- [Beb07] ———, *Why finite element discretizations can be factored by triangular hierarchical matrices*, SIAM J. Numer. Anal. **45** (2007), no. 4, 1472–1494.
- [BH03] M. Bebendorf and W. Hackbusch, *Existence of \mathcal{H} -matrix approximants to the inverse FE-matrix of elliptic operators with L^∞ -coefficients*, Numer. Math. **95** (2003), no. 1, 1–28.

- [BKP79] I. Babuška, R.B. Kellogg, and J. Pitkäranta, *Direct and inverse error estimates for finite elements with mesh refinements*, Numer. Math. **33** (1979), 447–471.
- [Bör10] S. Börm, *Approximation of solution operators of elliptic partial differential equations by \mathcal{H} - and \mathcal{H}^2 -matrices*, Numer. Math. **115** (2010), no. 2, 165–193.
- [Cia78] P.G. Ciarlet, *The finite element method for elliptic problems*, North-Holland Publishing Co., Amsterdam-New York-Oxford, 1978, Studies in Mathematics and its Applications, Vol. 4.
- [Clé75] Ph. Clément, *Approximation by finite element functions using local regularization*, Rev. Française Automat. Informat. Recherche Opérationnelle Sér. **9** (1975), no. R-2, 77–84.
- [DFG+01] W. Dahmen, B. Faermann, I. G. Graham, W. Hackbusch, and S. A. Sauter, *Inverse inequalities on non-quasiuniform meshes and application to the mortar element method*, Math. Comp. **73** (2001), 1107–1138.
- [FMP15] M. Faustmann, J.M. Melenk, and D. Praetorius, *H-matrix approximability of the inverses of FEM matrices*, Numer. Math. **131** (2015), no. 4, 615–642.
- [FMP16] M. Faustmann, J. M. Melenk, and D. Praetorius, *Existence of \mathcal{H} -matrix approximants to the inverse of BEM matrices: the simple-layer operator*, Math. Comp. **85** (2016), 119–152.
- [FMP17] ———, *Existence of \mathcal{H} -matrix approximants to the inverse of BEM matrices: the hyper-singular integral operator*, IMA J. Numer. Anal. **37** (2017), no. 3, 1211–1244.
- [GH03] L. Grasedyck and W. Hackbusch, *Construction and arithmetics of \mathcal{H} -matrices*, Computing **70** (2003), no. 4, 295–334.
- [GHK08] L. Grasedyck, W. Hackbusch, and R. Kriemann, *Performance of \mathcal{H} -LU preconditioning for sparse matrices*, Comput. Methods Appl. Math. **8** (2008), no. 4, 336–349.
- [GHLB04] L. Grasedyck, W. Hackbusch, and S. Le Borne, *Adaptive geometrically balanced clustering of h-matrices*, Computing **73** (2004), no. 1, 1–23.
- [GKLB08] L. Grasedyck, R. Kriemann, and S. Le Borne, *Parallel black box \mathcal{H} -LU preconditioning for elliptic boundary value problems*, Comput. Vis. Sci. **11** (2008), no. 4-6, 273–291.
- [GR97] L. Greengard and V. Rokhlin, *A new version of the fast multipole method for the Laplace in three dimensions*, Acta Numerica 1997, Cambridge University Press, 1997, pp. 229–269.
- [Gra01] L. Grasedyck, *Theorie und Anwendungen Hierarchischer Matrizen*, Ph.D. thesis, Universität Kiel, 2001.
- [Hac99] W. Hackbusch, *A sparse matrix arithmetic based on \mathcal{H} -matrices. Introduction to \mathcal{H} -matrices*, Computing **62** (1999), no. 2, 89–108.
- [Hac15] ———, *Hierarchical matrices: algorithms and analysis*, Springer Series in Computational Mathematics, vol. 49, Springer, Heidelberg, 2015.
- [Rok85] V. Rokhlin, *Rapid solution of integral equations of classical potential theory*, J. Comput. Phys. **60** (1985), 187–207.
- [Sch98] R. Schneider, *Multiskalen- und Wavelet-Matrixkompression: Analysisbasierte Methoden zur effizienten Lösung großer vollbesetzter Gleichungssysteme*, Advances in Numerical Mathematics, Teubner, 1998.
- [TW03] J. Tausch and J. White, *Multiscale bases for the sparse representation of boundary integral operators on complex geometry*, SIAM J. Sci. Comput. **24** (2003), no. 5, 1610–1629.
- [vPSS97] T. von Petersdorff, Ch. Schwab, and R. Schneider, *Multiwavelets for second-kind integral equations*, SIAM J. Numer. Anal. **34** (1997), no. 6, 2212–2227.