Numer. Math. (2015) 131:615–642
DOI 10.1007/s00211-015-0706-9

Numerische
Mathematik

$\mathcal{H}$-matrix approximability of the inverses of FEM matrices

Markus Faustmann · Jens Markus Melenk · Dirk Praetorius

Received: 2 August 2013 / Accepted: 23 November 2014 / Published online: 23 January 2015

Abstract We study the question of approximability for the inverse of the FEM stiffness matrix for (scalar) second order elliptic boundary value problems by blockwise low rank matrices such as those given by the $\mathcal{H}$-matrix format introduced by Hackbusch (Computing 62(2):89–108, 1999). We show that exponential convergence in the local block rank r can be achieved. We also show that exponentially accurate LU-decompositions in the $\mathcal{H}$-matrix format are possible for the stiffness matrices arising in the FEM. Our analysis avoids any coupling of the block rank r to the mesh width h. We also cover fairly general boundary conditions of mixed Dirichlet–Neumann–Robin boundary conditions.

Mathematics Subject Classification 65F05 · 65N30 · 65F30 · 65F50

1 Introduction

The format of $\mathcal{H}$-matrices was introduced in [27] as blockwise low-rank matrices that permit storage, application, and even a full (approximate) arithmetic with log-linear complexity, [20,22,28]. This data-sparse format is well suited to represent at high accuracy matrices arising as discretizations of many integral operators, for example, those appearing in boundary integral equation methods. Also the sparse matrices that

M. Faustmann (✉) · J. M. Melenk · D. Praetorius
Institute for Analysis and Scientific Computing (Inst. E 101), Vienna University of Technology,
Wiedner Hauptstrae 8-10, 1040 Vienna, Austria
e-mail: markus.faustmann@tuwien.ac.at

J. M. Melenk
e-mail: melenk@tuwien.ac.at

D. Praetorius
e-mail: dirk.praetorius@tuwien.ac.at

are obtained when discretizing differential operator by means of the finite element method (FEM) are amenable to a treatment by $\mathcal{H}$-matrices; in fact, they feature a lossless representation. Since the $\mathcal{H}$-matrix format comes with an arithmetic that provides algorithms to invert matrices as well as to compute LU-factorizations, approximations of the inverses of FEM matrices or their LU-factorizations are available computationally. Immediately, the question of accuracy and/or complexity comes into sight. On the one hand, the complexity of the $\mathcal{H}$-matrix inversion can be log-linear if the $\mathcal{H}$-matrix structure including the block ranks is fixed [20,22,28]. Then, however, the accuracy of the resulting approximate inverse is not completely clear. On the other hand, the accuracy of the inverse can be controlled by means of an adaptive arithmetic (going back at least to [20]); the computational cost at which this error control comes, is problem-dependent and not completely clear. Therefore, a fundamental question is how well the inverse can be approximated in a selected $\mathcal{H}$-matrix format, irrespective of algorithmic considerations. This question is answered in the present paper for FEM matrices arising from the discretization of second order elliptic boundary value problems.

It was first observed numerically in [20] that the inverse of the finite element (FEM) stiffness matrix corresponding to the Dirichlet problem for elliptic operators with bounded coefficients can be approximated in the format of $\mathcal{H}$-matrices with an error that decays exponentially in the block rank employed. Using properties of the continuous Green's function for the Dirichlet problem [4] proves this exponential decay in the block rank, at least up to the discretization error. The work [6] improves on the result [4] in several ways, in particular, by proving a corresponding approximation result in the framework of $\mathcal{H}^2$-matrices; we do not go into the details of $\mathcal{H}^2$-matrices here and merely mention that $\mathcal{H}^2$-matrices are a refinement of the concept of $\mathcal{H}$-matrices with better complexity properties, [7,18,29,30].

Whereas the analysis of [4,6] is based on the solution operator on the continuous level (i.e., by studying the Green's function), the approach taken in the present article is to work on the discrete level. This seemingly technical difference has several important ramifications: First, the exponential approximability in the block rank shown here is not limited by the discretization error as in [4,6]. Second, in contrast to [4,6], where the block rank r and the mesh width h are coupled by $r \sim |\log h|$, our estimates are explicit in both r and h. Third, a unified treatment of a variety of boundary conditions is possible and indeed worked out by us. Fourth, our approach paves the way for a similar approximability result for discretizations of boundary integral operators, [16,17]. Additionally, we mention that we also allow here the case of higher order FEM discretizations.

The last theoretical part of this paper (Sect. 5) shows that the $\mathcal{H}$-matrix format admits $\mathcal{H}$-LU-decompositions or $\mathcal{H}$-Cholesky factorizations with exponential accuracy in the block rank. This is achieved, following [3,11], by exploiting that the off-diagonal blocks of certain Schur complements are low-rank. Such an approach is closely related to the concepts of hierarchically semiseparable matrices (see, for example, [36,42,43] and references therein) and recursive skeletonization (see [26,32]) and their arithmetic. In fact, several multilevel "direct" solvers for PDE discretizations have been proposed in the recent past, [19,31,38,39]. These solvers take the form of (approximate) matrix factorizations. A key ingredient to their efficiency is that certain Schur complement blocks are compressible since they are low-rank. Thus, our analysis in Sect. 5 could also be of value for the understanding of these algorithms. We close by stressing that our analysis in Sect. 5 of $\mathcal{H}$-LU-decompositions makes very few assumptions on the actual ordering of the unknowns and does not explore beneficial features of special orderings. It is well-known in the context of classical direct solvers that the ordering of the unknowns has a tremendous impact on the fill-in in factorizations. One of the most successful techniques for discretizations of PDEs are multilevel nested dissection strategies, which permit to identify large matrix blocks that will not be filled during the factorization. An in-depth complexity analysis for the $\mathcal{H}$-matrix arithmetic for such ordering strategies can be found in [25]. The recent works [19,31] and, in a slightly different context, [5], owe at least parts of their efficiency to the use of nested dissection techniques.

2 Main results

Let $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$, be a bounded polygonal (for $d = 2$) or polyhedral (for $d = 3$) Lipschitz domain with boundary $\Gamma := \partial\Omega$. We consider differential operators of the form

$$Lu := -\operatorname{div}(\mathbf{C}\nabla u) + \mathbf{b} \cdot \nabla u + \beta u, \tag{1}$$

where $\mathbf{b} \in L^\infty(\Omega; \mathbb{R}^d)$, $\beta \in L^\infty(\Omega)$, and $\mathbf{C} \in L^\infty(\Omega; \mathbb{R}^{d\times d})$ is pointwise symmetric with

$$c_1 \|y\|_2^2 \leq \langle \mathbf{C}(x)y, y\rangle_2 \leq c_2 \|y\|_2^2 \quad \forall y \in \mathbb{R}^d, \tag{2}$$

with certain constants $c_1, c_2 > 0$.

For $f \in L^2(\Omega)$, we consider the mixed boundary value problem

$$\begin{aligned} Lu &= f \quad \text{in } \Omega, && \text{(3a)}\\ u &= 0 \quad \text{on } \Gamma_D, && \text{(3b)}\\ \mathbf{C}\nabla u \cdot n &= 0 \quad \text{on } \Gamma_N, && \text{(3c)}\\ \mathbf{C}\nabla u \cdot n + \alpha u &= 0 \quad \text{on } \Gamma_{\mathcal{R}}, && \text{(3d)} \end{aligned}$$

where n denotes the outer normal vector to the surface Γ, $\alpha \in L^\infty(\Gamma_{\mathcal{R}})$, $\alpha > 0$ and $\Gamma = \overline{\Gamma_D} \cup \overline{\Gamma_N} \cup \overline{\Gamma_{\mathcal{R}}}$, with the pairwise disjoint and relatively open subsets $\Gamma_D, \Gamma_N, \Gamma_{\mathcal{R}}$. With the trace operator γ_0^{int} we define $H_0^1(\Omega, \Gamma_D) := \{u \in H^1(\Omega) : \gamma_0^{\text{int}} u = 0 \text{ on } \Gamma_D\}$. The bilinear form $a : H_0^1(\Omega, \Gamma_D) \times H_0^1(\Omega, \Gamma_D) \to \mathbb{R}$ corresponding to (3) is given by

$$a(u, v) := \langle \mathbf{C}\nabla u, \nabla v\rangle_{L^2(\Omega)} + \langle \mathbf{b} \cdot \nabla u + \beta u, v\rangle_{L^2(\Omega)} + \langle \alpha u, v\rangle_{L^2(\Gamma_{\mathcal{R}})}. \tag{4}$$

We additionally assume that the coefficients $\alpha, \mathbf{C}, \mathbf{b}, \beta$ are such that the coercivity

$$\|u\|_{H^1(\Omega)}^2 \leq Ca(u, u) \tag{5}$$

of the bilinear form $a(\cdot, \cdot)$ holds. Then, the Lax–Milgram Lemma implies the unique solvability of the weak formulation of our model problem.

For the discretization, we assume that Ω is triangulated by a *quasiuniform* mesh $\mathcal{T}_h = \{T_1, \dots, T_N\}$ of mesh width $h := \max_{T_j \in \mathcal{T}_h} \operatorname{diam}(T_j)$, and the Dirichlet Γ_D, Neumann Γ_N, and Robin $\Gamma_{\mathcal{R}}$-parts of the boundary are resolved by the mesh $\mathcal{T}_h$. The elements $T_j \in \mathcal{T}_h$ are triangles ($d = 2$) or tetrahedra ($d = 3$), and we assume that $\mathcal{T}_h$ is regular in the sense of Ciarlet. The nodes are denoted by $x_i \in \mathcal{N}_h$, for $i = 1, \dots, N$. Moreover, the mesh $\mathcal{T}_h$ is assumed to be γ-shape regular in the sense of $h \sim \operatorname{diam}(T_j) \le \gamma\, |T_j|^{1/d}$ for all $T_j \in \mathcal{T}_h$. In the following, the notation $\lesssim$ abbreviates $\le$ up to a constant $C > 0$ which depends only on Ω, the dimension d, and γ-shape regularity of $\mathcal{T}_h$. Moreover, we use $\simeq$ to abbreviate that both estimates $\lesssim$ and $\gtrsim$ hold.

We consider the Galerkin discretization of the bilinear form $a(\cdot,\cdot)$ by continuous, piecewise polynomials of fixed degree $p \ge 1$ in $S_0^{p,1}(\mathcal{T}_h, \Gamma_D) := S^{p,1}(\mathcal{T}_h) \cap H_0^1(\Omega, \Gamma_D)$ with $S^{p,1}(\mathcal{T}_h) = \{u \in C(\Omega) : u|_{T_j} \in \mathcal{P}_p, \forall T_j \in \mathcal{T}_h\}$, where $\mathcal{P}_p$ denotes the space of polynomials of degree p. We choose a basis of $S_0^{p,1}(\mathcal{T}_h, \Gamma_D)$, which is denoted by $\mathcal{B}_h := \{\psi_j : j = 1, \dots, N\}$. Given that our results are formulated for matrices, assumptions on the basis $\mathcal{B}_h$ need to be imposed. For the isomorphism $\mathcal{J} : \mathbb{R}^N \to S_0^{p,1}(\mathcal{T}_h, \Gamma_D)$, $\mathbf{x} \mapsto \sum_{j=1}^N x_j \psi_j$, we require

$$h^{d/2}\, \|\mathbf{x}\|_2 \lesssim \|\mathcal{J}\mathbf{x}\|_{L^2(\Omega)} \lesssim h^{d/2}\, \|\mathbf{x}\|_2\,, \quad \forall \mathbf{x} \in \mathbb{R}^d. \tag{6}$$

Remark 1 Standard bases for $p = 1$ are the classical hat functions satisfying $\psi_j(x_i) = \delta_{ij}$ and for $p \ge 2$ we refer to, e.g., [13,34,40].

The Galerkin discretization of (4) results in a positive definite matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$ with

$$\begin{aligned}\mathbf{A}_{jk} = \left\langle \mathbf{C}\nabla\psi_k, \nabla\psi_j\right\rangle_{L^2(\Omega)} + \left\langle \mathbf{b}\cdot\nabla\psi_k + \beta\psi_k, \psi_j\right\rangle_{L^2(\Omega)} \\ + \left\langle \alpha\psi_k, \psi_j\right\rangle_{L^2(\Gamma_{\mathcal{R}})}, \quad \psi_k, \psi_j \in \mathcal{B}_h.\end{aligned} \tag{7}$$

Our goal is to derive an $\mathcal{H}$-matrix approximation $\mathbf{B}_{\mathcal{H}}$ of the inverse matrix $\mathbf{B} = \mathbf{A}^{-1}$. An $\mathcal{H}$-matrix $\mathbf{B}_{\mathcal{H}}$ is a blockwise low rank matrix based on the concept of “admissibility”, which we now introduce:

Definition 1 (*Bounding boxes and η-admissibility*) A *cluster* τ is a subset of the index set $\mathcal{I} = \{1, \dots, N\}$. For a cluster $\tau \subset \mathcal{I}$, we say that $B_{R_\tau} \subset \mathbb{R}^d$ is a *bounding box* if:

(i) B_{R_τ} is a hyper cube with side length R_τ,
(ii) $\operatorname{supp}\psi_j \subset B_{R_\tau}$ for all $j \in \tau$.

For $\eta > 0$, a pair of clusters (τ, σ) with $\tau, \sigma \subset \mathcal{I}$ is *η-admissible*, if there exist boxes B_{R_τ}, B_{R_σ} satisfying (i)–(ii) such that

$$\max\{\operatorname{diam} B_{R_\tau}, \operatorname{diam} B_{R_\sigma}\} \le \eta\, \operatorname{dist}(B_{R_\tau}, B_{R_\sigma}). \tag{8}$$

Definition 2 (*blockwise rank-r matrices*) Let P be a partition of $\mathcal{I} \times \mathcal{I}$ and $\eta > 0$ an admissibility parameter. A matrix $\mathbf{B}_{\mathcal{H}} \in \mathbb{R}^{N\times N}$ is said to be a *blockwise rank-r matrix*, if for every η-admissible cluster pair $(\tau, \sigma) \in P$, the block $\mathbf{B}_{\mathcal{H}}|_{\tau\times\sigma}$ is a rank-r

matrix, i.e., it has the form $\mathbf{B}_{\mathcal{H}}|_{\tau\times\sigma} = \mathbf{X}_{\tau\sigma}\mathbf{Y}_{\tau\sigma}^T$ with $\mathbf{X}_{\tau\sigma} \in \mathbb{R}^{|\tau|\times r}$ and $\mathbf{Y}_{\tau\sigma} \in \mathbb{R}^{|\sigma|\times r}$. Here and below, $|\sigma|$ denotes the cardinality of a finite set σ, and $\mathbf{B}|_{\tau\times\sigma} \in \mathbb{R}^{|\tau|\times|\sigma|}$ denotes the $\tau \times \sigma$ subblock of a matrix $\mathbf{B}$ as a $|\tau| \times |\sigma|$-matrix.

The following theorems are the main results of this paper. Theorem 1 shows that admissible blocks can be approximated by rank-r matrices:

Theorem 1 *Fix the admissibility parameter $\eta > 0$, $q \in (0, 1)$. Let the cluster pair (τ, σ) be η-admissible. Then, for $k \in \mathbb{N}$ there are matrices $\mathbf{X}_{\tau\sigma} \in \mathbb{R}^{|\tau|\times r}$, $\mathbf{Y}_{\tau\sigma} \in \mathbb{R}^{|\sigma|\times r}$ of rank $r \leq C_{\text{dim}}(2+\eta)^d q^{-d} k^{d+1}$ such that*

$$\left\| \mathbf{A}^{-1}|_{\tau\times\sigma} - \mathbf{X}_{\tau\sigma}\mathbf{Y}_{\tau\sigma}^T \right\|_2 \leq C_{\text{apx}} N q^k. \tag{9}$$

The constants $C_{\text{apx}}, C_{\text{dim}} > 0$ depend only on the boundary value problem (3), Ω, d, p, and the γ-shape regularity of the quasiuniform triangulation $\mathcal{T}_h$.

The approximations for the individual blocks can be combined to gauge the approximability of $\mathbf{A}^{-1}$ by blockwise rank-r matrices. Particularly satisfactory estimates are obtained if the blockwise rank-r matrices have additional structure. To that end, we introduce the following definitions.

Definition 3 (*Cluster tree*) A *cluster tree* with *leaf size* $n_{\text{leaf}} \in \mathbb{N}$ is a binary tree $\mathbb{T}_{\mathcal{I}}$ with root $\mathcal{I}$ such that for each cluster $\tau \in \mathbb{T}_{\mathcal{I}}$ the following dichotomy holds: either τ is a leaf of the tree and $|\tau| \leq n_{\text{leaf}}$, or there exist sons $\tau', \tau'' \in \mathbb{T}_{\mathcal{I}}$, which are disjoint subsets of τ with $\tau = \tau' \cup \tau''$. The *level function* level : $\mathbb{T}_{\mathcal{I}} \to \mathbb{N}_0$ is inductively defined by $\text{level}(\mathcal{I}) = 0$ and $\text{level}(\tau') := \text{level}(\tau) + 1$ for τ' a son of τ. The *depth* of a cluster tree is $\text{depth}(\mathbb{T}_{\mathcal{I}}) := \max_{\tau\in\mathbb{T}_{\mathcal{I}}} \text{level}(\tau)$.

Definition 4 (*Far field, near field, and sparsity constant*) A partition P of $\mathcal{I} \times \mathcal{I}$ is said to be based on the cluster tree $\mathbb{T}_{\mathcal{I}}$, if $P \subset \mathbb{T}_{\mathcal{I}} \times \mathbb{T}_{\mathcal{I}}$. For such a partition P and fixed admissibility parameter $\eta > 0$, we define the *far field* and the *near field* as

$$P_{\text{far}} := \{(\tau, \sigma) \in P \ : \ (\tau, \sigma) \ \text{ is } \eta\text{-admissible } \}, \quad P_{\text{near}} := P \backslash P_{\text{far}}.$$

The *sparsity constant* C_{sp}, introduced in [20], of such a partition is defined by

$$C_{\text{sp}} := \max\left\{ \max_{\tau\in\mathbb{T}_{\mathcal{I}}} |\{\sigma \in \mathbb{T}_{\mathcal{I}} \ : \ \tau \times \sigma \in P_{\text{far}}\}|\, , \max_{\sigma\in\mathbb{T}_{\mathcal{I}}} |\{\tau \in \mathbb{T}_{\mathcal{I}} \ : \ \tau \times \sigma \in P_{\text{far}}\}| \right\}.$$

The following Theorem 2 shows that the matrix $\mathbf{A}^{-1}$ can be approximated by blockwise rank-r matrices at an exponential rate in the block rank r:

Theorem 2 *Fix the admissibility parameter $\eta > 0$. Let a partition P of $\mathcal{I} \times \mathcal{I}$ be based on a cluster tree $\mathbb{T}_{\mathcal{I}}$. Then, there is a blockwise rank-r matrix $\mathbf{B}_{\mathcal{H}}$ such that*

$$\left\| \mathbf{A}^{-1} - \mathbf{B}_{\mathcal{H}} \right\|_2 \leq C_{\text{apx}} C_{\text{sp}} N \text{depth}(\mathbb{T}_{\mathcal{I}}) e^{-br^{1/(d+1)}}. \tag{10}$$

The constants $C_{\text{apx}}, b > 0$ depend only on the boundary value problem (3), Ω, d, p, and the γ-shape regularity of the quasiuniform triangulation $\mathcal{T}_h$.

Remark 2 Typical clustering strategies such as the "geometric clustering" described in [28] and applied to quasiuniform meshes with $\mathcal{O}(N)$ elements lead to fairly balanced cluster trees $\mathbb{T}_{\mathcal{I}}$ of depth $\mathcal{O}(\log N)$ and feature a sparsity constant $C_{\rm sp}$ that is bounded uniformly in N. We refer to [28] for the fact that the memory requirement to store $\mathbf{B}_{\mathcal{H}}$ is $\mathcal{O}((r + n_{\rm leaf})N \log N)$.

Remark 3 With the estimate $\frac{1}{\|\mathbf{A}^{-1}\|_2} \lesssim N^{-1}$ from [14, Theorem 2], we get a bound for the relative error

$$\frac{\|\mathbf{A}^{-1} - \mathbf{B}_{\mathcal{H}}\|_2}{\|\mathbf{A}^{-1}\|_2} \lesssim C_{\rm apx} C_{\rm sp} \operatorname{depth}(\mathbb{T}_{\mathcal{I}}) e^{-br^{1/(d+1)}}. \tag{11}$$

Let us conclude this section with an observation concerning the admissibility condition (8). If the operator L is symmetric, i.e., $\mathbf{b} = \mathbf{0}$, then the admissibility condition (8) can be replaced by the weaker admissibility condition

$$\min\{\operatorname{diam} B_{R_\tau}, \operatorname{diam} B_{R_\sigma}\} \leq \eta \operatorname{dist}(B_{R_\tau}, B_{R_\sigma}). \tag{12}$$

This follows from the fact that Proposition 1 only needs an admissibility criterion of the form $\operatorname{diam} B_{R_\tau} \leq \eta \operatorname{dist}(B_{R_\tau}, B_{R_\sigma})$. Due to the symmetry of L, deriving a block approximation for the block $\tau \times \sigma$ is equivalent to deriving an approximation for the block $\sigma \times \tau$. Therefore, we can interchange roles of the boxes B_{R_τ} and B_{R_σ}, and as a consequence the weaker admissibility condition (12) is sufficient. We summarize this observation in the following corollary.

Corollary 1 *In the symmetric case* $\mathbf{b} = \mathbf{0}$*, the results from Theorems 1 and 2 hold verbatim with the weaker admissibility criterion* (12) *instead of* (8).

3 Low-dimensional approximation of the Galerkin solution on admissible blocks

In terms of functions and function spaces, the question of approximating the matrix block $\mathbf{A}^{-1}|_{\tau\times\sigma}$ by a low-rank factorization $\mathbf{X}_{\tau\sigma}\mathbf{Y}_{\tau\sigma}^T$ can be rephrased as one of how well one can approximate locally the solution of certain variational problems. More precisely, we consider, for data f supported by $B_{R_\sigma} \cap \Omega$, the problem to find $\phi_h \in S_0^{p,1}(\mathcal{T}_h, \Gamma_D)$ such that

$$a(\phi_h, \psi_h) = \langle f, \psi_h \rangle_{L^2(\Omega)}, \qquad \forall \psi_h \in S_0^{p,1}(\mathcal{T}_h, \Gamma_D). \tag{13}$$

We remark in passing that existence and uniqueness of ϕ_h follow from coercivity of $a(\cdot,\cdot)$. The question of approximating the matrix block $\mathbf{A}^{-1}|_{\tau\times\sigma}$ by a low-rank factorization is intimately linked to the question of approximating $\phi_h|_{B_{R_\tau}\cap\Omega}$ from low-dimensional spaces. The latter problem is settled in the affirmative in the following proposition for η-admissible cluster pairs (τ, σ):

Proposition 1 *Let* (τ, σ) *be a cluster pair with bounding boxes* B_{R_τ}, B_{R_σ}. *Assume that* $\eta \operatorname{dist}(B_{R_\tau}, B_{R_\sigma}) \geq \operatorname{diam}(B_{R_\tau})$ *for a fixed admissibility parameter* $\eta > 0$. *Fix* $q \in (0,1)$. *Let* $\Pi^{L^2} : L^2(\Omega) \to S_0^{p,1}(\mathcal{T}_h, \Gamma_D)$ *be the* $L^2(\Omega)$*-orthogonal projection. Then, for each* $k \in \mathbb{N}$ *there exists a space* $V_k \subset S_0^{p,1}(\mathcal{T}_h, \Gamma_D)$ *with* $\dim V_k \leq C_{\text{dim}}(2+\eta)^d q^{-d} k^{d+1}$ *such that for arbitrary* $f \in L^2(\Omega)$ *with* $\operatorname{supp} f \subset B_{R_\sigma} \cap \Omega \subset F_\tau := \{x \in \Omega : \eta \operatorname{dist}(x, B_{R_\tau}) \geq \operatorname{diam}(B_{R_\tau})\}$, *the solution* ϕ_h *of* (13) *satisfies*

$$\min_{v \in V_k} \|\phi_h - v\|_{L^2(B_{R_\tau} \cap \Omega)} \leq C_{\text{box}} q^k \|\Pi^{L^2} f\|_{L^2(\Omega)} \leq C_{\text{box}} q^k \|f\|_{L^2(B_{R_\sigma} \cap \Omega)}. \tag{14}$$

The constant $C_{\text{box}} > 0$ *depends only on the boundary value problem (3) and* Ω, *while* $C_{\text{dim}} > 0$ *additionally depends on* p, d, *and the* γ*-shape regularity of the quasiuniform triangulation* $\mathcal{T}_h$.

The proof of Proposition 1 will be given at the end of this section. The basic steps are as follows: First, one observes that $\operatorname{supp} f \subset B_{R_\sigma} \cap \Omega$ together with the admissibility condition $\operatorname{dist}(B_{R_\tau}, B_{R_\sigma}) \geq \eta^{-1} \operatorname{diam}(B_{R_\tau}) > 0$ imply the orthogonality condition

$$a(\phi_h, \psi_h) = \langle f, \psi_h \rangle_{L^2(B_{R_\sigma} \cap \Omega)} = 0, \quad \forall \psi_h \in S_0^{p,1}(\mathcal{T}_h, \Gamma_D) \text{ with } \operatorname{supp} \psi_h \subset B_{R_\tau} \cap \Omega. \tag{15}$$

Second, this observation will allow us to prove a Caccioppoli-type estimate (Lemma 2) in which stronger norms of ϕ_h are estimated by weaker norms of ϕ_h on slightly enlarged regions. Third, we proceed as in [4,6] by iterating an approximation result (Lemma 3) derived from Scott–Zhang interpolation of the Galerkin solution ϕ_h. This iteration argument accounts for the exponential convergence (Lemma 4).

3.1 The space $\mathcal{H}_h(D, \omega)$ and a Caccioppoli type estimate

It will be convenient to introduce, for index sets $\rho \subset \mathcal{I}$, the set

$$\omega_\rho := \text{interior}\left(\bigcup_{j \in \rho} \operatorname{supp} \psi_j\right) \subseteq \Omega; \tag{16}$$

we will implicitly assume henceforth that such sets are unions of elements. Let $D \subset \mathbb{R}^d$ be a bounded open set and $\omega \subset \Omega$ be of the form (16). The orthogonality property that we have identified in (15) is captured by the following space $\mathcal{H}_h(D, \omega)$:

$$\mathcal{H}_h(D, \omega) := \Big\{ u \in H^1(D \cap \omega) : \exists \widetilde{u} \in S_0^{p,1}(\mathcal{T}_h, \Gamma_D) \text{ s.t. } u|_{D \cap \omega} = \widetilde{u}|_{D \cap \omega}, \operatorname{supp} \widetilde{u} \subset \overline{\omega}, \\ a(u, \psi_h) = 0, \ \forall \ \psi_h \in S_0^{p,1}(\mathcal{T}_h, \Gamma_D) \text{ with } \operatorname{supp} \psi_h \subset \overline{D \cap \omega} \Big\}. \tag{17}$$

For the proof of Proposition 1 and subsequently Theorems 1 and 2, we will only need the special case $\omega = \Omega$; the general case $\mathcal{H}_h(D, \omega)$ with $\omega \neq \Omega$ will be required in our analysis of LU-decompositions in Sect. 5.2.

Clearly, the finite dimensional space $\mathcal{H}_h(D,\omega)$ is a closed subspace of $H^1(D\cap\omega)$, and we have $\phi_h \in \mathcal{H}_h(B_{R_\tau},\Omega)$ for the solution ϕ_h of (13) with supp $f \subset B_{R_\sigma}\cap\Omega$ and bounding boxes B_{R_τ}, B_{R_σ} that satisfy the η-admissibility criterion (8). Since multiplications of elements of $\mathcal{H}_h(D,\omega)$ with cut-off function and trivial extensions to Ω appear repeatedly in the sequel, we note the following very simple lemma:

Lemma 1 *Let ω be a union of elements, $D\subset\mathbb{R}^d$ be bounded and open, and $\eta\in W^{1,\infty}(\mathbb{R}^d)$ with* $\operatorname{supp}\eta\subset D$. *For $u\in\mathcal{H}_h(D,\omega)$ define the function ηu pointwise on Ω by $(\eta u)(x):=\eta(x)u(x)$ for $x\in D\cap\omega$ and $(\eta u)(x)=0$ for $x\notin D\cap\omega$. Then*

(i) $\eta u\in H^1_0(\Omega;\Gamma_D)$
(ii) $\operatorname{supp}(\eta u)\subset\overline{D\cap\omega}$
(iii) *If $\eta\in S^{q,1}(\mathcal{T}_h)$, then $\eta u\in S_0^{p+q,1}(\mathcal{T}_h,\Gamma_D)$.*

Proof We only illustrate (i). Given $u\in\mathcal{H}_h(D,\omega)$ there exists by definition a function $\widetilde{u}\in S_0^{p,1}(\mathcal{T}_h,\Gamma_D)$ with $\operatorname{supp}\widetilde{u}\subset\overline{\omega}$. By the support properties of η and $\widetilde{u}$, the function ηu coincides with $\eta\widetilde{u}$. As the product of an $H^1(\Omega)$-function and a Lipschitz continuous function, the function $\eta\widetilde{u}$ is in $H^1(\Omega)$.

A main tool in our proofs is the nodal interpolation operator $J_h : C(\overline{\Omega})\to S_0^{p,1}(\mathcal{T}_h,\Gamma_D)$. Since $p+1>\frac{d}{2}$, the interpolation operator J_h has the following local approximation property for continuous, $\mathcal{T}_h$-piecewise H^{p+1}-functions $u\in C(\overline{\Omega})\cap H^{p+1}_{\mathrm{pw}}(\mathcal{T}_h,\omega):=\{u\in L^2(\omega): u|_T\in H^{p+1}(T)\,\forall T\in\mathcal{T}_h\}$

$$\|u-J_hu\|^2_{H^m(T)}\le Ch^{2(p+1-m)}\,|u|^2_{H^{p+1}(T)},\quad 0\le m\le p+1. \tag{18}$$

The constant $C>0$ depends only on γ-shape regularity of the quasiuniform triangulation $\mathcal{T}_h$, the dimension d, and the polynomial degree p. In particular, it is independent of the choice of the set ω.

In the following, we will construct approximations on nested boxes and therefore introduce the notion of concentric boxes.

Definition 5 (*Concentric boxes*) Two boxes $B_R, B_{R'}$ of side length R, R' are said to be concentric, if they have the same barycenter and B_R can be obtained by a stretching of $B_{R'}$ by the factor R/R' taking their common barycenter as the origin.

For a box B_R with side length $R\le 2\operatorname{diam}(\Omega)$, we introduce the norm

$$|||u|||^2_{h,R}:=\left(\frac{h}{R}\right)^2\|\nabla u\|^2_{L^2(B_R\cap\omega)}+\frac{1}{R^2}\|u\|^2_{L^2(B_R\cap\omega)},$$

which is, for fixed h, equivalent to the H^1-norm.

The following lemma states a discrete Caccioppoli-type estimate for functions in $\mathcal{H}_h(B_{(1+\delta)R},\omega)$, where $B_{(1+\delta)R}$ and B_R are concentric boxes. In contrast to the classical, continuous Caccioppoli inequality, an additional assumption on the size parameters R,δ of the box $B_{(1+\delta)R}$ compared to the mesh size h has to be made.

Lemma 2 *Let $\delta \in (0,1)$, $R \in (0, 2\operatorname{diam}(\Omega))$ such that $\frac{h}{R} \le \frac{\delta}{4}$ and let $\omega \subseteq \Omega$ be of the form (16). Let B_R, $B_{(1+\delta)R}$ be two concentric boxes. Let $u \in \mathcal{H}_h(B_{(1+\delta)R}, \omega)$. Then, there exists a constant $C_{reg} > 0$, which depends only on the boundary value problem (3), Ω, d, p, and the γ-shape regularity of the quasiuniform triangulation $\mathcal{T}_h$, such that*

$$\|\nabla u\|_{L^2(B_R\cap\omega)} \le \|\nabla u\|_{L^2(B_R\cap\omega)} + \langle \alpha u, u\rangle^{1/2}_{L^2(B_R\cap(\Gamma_{\mathcal{R}}\cap\overline{\omega}))} \le C_{reg}\frac{1+\delta}{\delta}\, |||u|||_{h,(1+\delta)R}\,. \tag{19}$$

Proof Let $\eta \in S^{1,1}(\mathcal{T}_h)$ be a piecewise affine cut-off function with $\operatorname{supp}\eta \subset B_{(1+\delta/2)R} \cap \overline{\Omega}$, $\eta \equiv 1$ on $B_R \cap \omega$, $0 \le \eta \le 1$, and $\|\nabla\eta\|_{L^\infty(B_{(1+\delta)R}\cap\Omega)} \lesssim \frac{1}{\delta R}$. By Lemma 1 we have $\eta^2 u \in S_0^{p+2,1}(\mathcal{T}_h, \Gamma_D) \subset H_0^1(\Omega; \Gamma_D)$ and

$$\operatorname{supp}(\eta^2 u) \subset \overline{B_{(1+\delta/2)R} \cap \omega}. \tag{20}$$

Recall that h is the maximal element diameter and $4h \le \delta R$. Hence, for the nodal interpolation operator J_h, we have $\operatorname{supp} J_h(\eta^2 u) \subset \overline{B_{(1+\delta)R}}$; in view of the locality of the nodal interpolation, we furthermore have $\operatorname{supp} J_h(\eta^2 u) \subset \overline{\omega}$ so that

$$\operatorname{supp} J_h(\eta^2 u) \subset \overline{B} \quad \text{with}\, B := B_{(1+\delta)R} \cap \omega. \tag{21}$$

With the coercivity of the bilinear form $a(\cdot,\cdot)$ and $\frac{1}{\delta R} \lesssim \frac{1}{\delta^2 R^2}$, since $\delta < 1$ and $R \le 2\operatorname{diam}(\Omega)$, we have

$$\|\nabla u\|^2_{L^2(B_R\cap\omega)} + \langle \alpha u, u\rangle_{L^2(\overline{B_R\cap\omega}\cap\Gamma_R)} \le \|\nabla(\eta u)\|^2_{L^2(B)} + \langle \alpha\eta u, \eta u\rangle_{L^2(\overline{B}\cap\Gamma_R)} \tag{22a}$$

$$\begin{aligned}
&\lesssim a(\eta u, \eta u)\\
&= \int_B \mathbf{C}\nabla u \cdot \nabla(\eta^2 u) + u^2\mathbf{C}\nabla\eta\cdot\nabla\eta\, dx + \left\langle \mathbf{b}\cdot\nabla u + \beta u, \eta^2 u\right\rangle_{L^2(B)}\\
&\quad + \langle \mathbf{b}\cdot(\nabla\eta)u, \eta u\rangle_{L^2(B)} + \left\langle \alpha u, \eta^2 u\right\rangle_{L^2(\overline{B}\cap\Gamma_{\mathcal{R}})}\\
&\lesssim \int_B \mathbf{C}\nabla u\cdot\nabla(\eta^2 u)dx + \left\langle \mathbf{b}\cdot\nabla u + \beta u, \eta^2 u\right\rangle_{L^2(B)}\\
&\quad + \left\langle \alpha u, \eta^2 u\right\rangle_{L^2(\overline{B}\cap\Gamma_{\mathcal{R}})} + \frac{1}{\delta^2 R^2}\|u\|^2_{L^2(B)}\\
&= a(u, \eta^2 u) + \frac{1}{\delta^2 R^2}\|u\|^2_{L^2(B)}\,.
\end{aligned} \tag{22b}$$

Recall from (21) that supp $J_h(\eta^2 u) \subset \overline{B}$. The orthogonality relation (17) in the definition of the space $\mathcal{H}_h(B, \omega)$ therefore implies

$$\begin{aligned} a(u, \eta^2 u) &= a(u, \eta^2 u - J_h(\eta^2 u)) \\ &\leq \|\mathbf{C}\|_{L^\infty(B)} \|\nabla u\|_{L^2(B)} \left\|\nabla(\eta^2 u - J_h(\eta^2 u))\right\|_{L^2(B)} \\ &\quad + \left(\|\mathbf{b}\|_{L^\infty(B)} \|\nabla u\|_{L^2(B)} + \|\beta\|_{L^\infty(B)} \|\eta u\|_{L^2(B)}\right) \left\|\eta^2 u - J_h(\eta^2 u)\right\|_{L^2(B)} \\ &\quad + \left|\left\langle \alpha u, \eta^2 u - J_h(\eta^2 u)\right\rangle_{L^2(\overline{B} \cap \Gamma_{\mathcal{R}})}\right|. \end{aligned} \tag{23}$$

The approximation property (18) and the support properties of $\eta^2 u$ lead to

$$\left\|\nabla(\eta^2 u - J_h(\eta^2 u))\right\|^2_{L^2(\Omega)} \lesssim h^{2p} \sum_{\substack{T \in \mathcal{T}_h \\ T \subseteq B}} \left|\eta^2 u\right|^2_{H^{p+1}(T)}. \tag{24}$$

By $D^k u$, we denote the derivative $D^k u := \frac{\partial^{|k|} u}{\partial x_1^{k_1} \cdots \partial x_d^{k_d}}$ with the multi-index $k \in \mathbb{N}_0^d$ with $|k| = \sum_{i=1}^d k_i$. Since, for each $T \subset B$ we have $u|_T \in \mathcal{P}_p$, we get $D^k u|_T = 0$ for all multi-indices $k \in \mathbb{N}_0^d$ with $|k| = p + 1$. The assumption $\eta \in S^{1,1}(\mathcal{T}_h)$ implies $D^k \eta|_T = 0$ for all $k \in \mathbb{N}_0^d$ with $|k| \geq 2$. With the Leibniz product rule, the right-hand side of (24) can therefore be estimated by

$$\begin{aligned} |\eta^2 u|^2_{H^{p+1}(T)} &= \sum_{\substack{k \in \mathbb{N}_0^d \\ |k| = p+1}} \left\|D^k(\eta^2 u)\right\|^2_{L^2(T)} \lesssim \sum_{\substack{k \in \mathbb{N}_0^d \\ |k| = p+1}} \left\| \sum_{\substack{\ell \in \mathbb{N}_0^d, \ell \leq k \\ |\ell| \leq 2}} \binom{k}{\ell} D^{k-\ell} u D^\ell(\eta^2) \right\|^2_{L^2(T)} \\ &\lesssim \sum_{\substack{k \in \mathbb{N}_0^d \\ |k| = p+1}} \left\| \sum_{\substack{\ell \in \mathbb{N}_0^d, \ell \leq k \\ |\ell| = 1}} \binom{k}{\ell} D^{k-\ell} u (D^\ell \eta)\eta + \sum_{\substack{m \in \mathbb{N}_0^d, m \leq k-\ell \\ |m| = 1}} \binom{k}{\ell + m} D^{k-\ell-m} u D^\ell \eta D^m \eta \right\|^2_{L^2(T)} \\ &\lesssim \sum_{\substack{k \in \mathbb{N}_0^d \\ |k| = p+1}} \sum_{\substack{\ell \in \mathbb{N}_0^d, \ell \leq k \\ |\ell| = 1}} \left|\binom{k}{\ell}\right|^2 \left\| D^{k-\ell} u (D^\ell \eta)\eta + D^\ell \eta \sum_{\substack{m \in \mathbb{N}_0^d, m \leq k-\ell \\ |m| = 1}} \binom{k}{\ell}^{-1} \binom{k}{\ell + m} D^{k-\ell-m} u D^m \eta \right\|^2_{L^2(T)} \\ &\lesssim \frac{1}{(\delta R)^2} \sum_{\substack{k \in \mathbb{N}_0^d \\ |k| = p+1}} \sum_{\substack{\ell \in \mathbb{N}_0^d, \ell \leq k \\ |\ell| = 1}} |k|^2 \left\| (D^{k-\ell} u)\eta + \sum_{\substack{m \in \mathbb{N}_0^d, m \leq k-\ell \\ |m| = 1}} \binom{k-\ell}{m} D^{k-\ell-m} u D^m \eta \right\|^2_{L^2(T)} \\ &\quad + \frac{1}{(\delta R)^2} \sum_{\substack{k \in \mathbb{N}_0^d \\ |k| = p+1}} \sum_{\substack{\ell \in \mathbb{N}_0^d, \ell \leq k \\ |\ell| = 1}} |k|^2 \left\| \sum_{\substack{m \in \mathbb{N}_0^d, m \leq k-\ell \\ |m| = 1}} \left(\binom{k}{\ell}^{-1} \binom{k}{\ell + m} - \binom{k-\ell}{m}\right) D^{k-\ell-m} u D^m \eta \right\|^2_{L^2(T)} \end{aligned}$$

$$\begin{aligned}
&\lesssim \frac{1}{(\delta R)^2} \sum_{\substack{k\in\mathbb{N}_0^d\\ |k|=p+1}} \sum_{\substack{\ell\in\mathbb{N}_0^d,\ell\le k\\ |\ell|=1}} \left\|D^{k-\ell}(u\eta)\right\|^2_{L^2(T)} + \frac{1}{(\delta R)^4}\left|u\right|^2_{H^{p-1}(T)}\\
&\lesssim \frac{1}{(\delta R)^2}\left|\eta u\right|^2_{H^p(T)} + \frac{1}{(\delta R)^4}\left|u\right|^2_{H^{p-1}(T)},
\end{aligned}$$

where the suppressed constant depends on p. The inverse inequality $|\eta u|_{H^p(T)} \lesssim h^{-p+1}\left\|\nabla(\eta u)\right\|_{L^2(T)}$, see, e.g., [12], leads to

$$\begin{aligned}
\left\|\nabla(\eta^2 u - J_h(\eta^2 u))\right\|^2_{L^2(\Omega)} &\lesssim \frac{1}{(\delta R)^2} h^{2p} \sum_{\substack{T\in\mathcal{T}_h\\ T\subseteq B}} \left(\left|\eta u\right|^2_{H^p(T)} + \frac{1}{(\delta R)^2}\left|u\right|^2_{H^{p-1}(T)}\right)\\
&\lesssim \frac{h^2}{(\delta R)^2}\left\|\nabla(\eta u)\right\|^2_{L^2(B)} + \frac{h^2}{(\delta R)^4}\left\|u\right\|^2_{L^2(B)}\\
&\lesssim \frac{h^2}{(\delta R)^4}\left\|u\right\|^2_{L^2(B)} + \frac{h^2}{(\delta R)^2}\left\|\eta\nabla u\right\|^2_{L^2(B)}. \qquad (25)
\end{aligned}$$

The same line of reasoning leads to

$$\left\|\eta^2 u - J_h(\eta^2 u)\right\|_{L^2(\Omega)} \lesssim \frac{h^2}{(\delta R)^2}\left\|u\right\|_{L^2(B)} + \frac{h^2}{\delta R}\left\|\eta\nabla u\right\|_{L^2(B)}. \qquad (26)$$

In order to derive an estimate for the boundary term in (23), we need a second smooth cut-off function $\widetilde{\eta}$ with $\operatorname{supp}\widetilde{\eta} \subset \overline{B_{(1+\delta)R}}$ and $\widetilde{\eta} \equiv 1$ on $\operatorname{supp}(J_h(\eta^2 u) - \eta^2 u)$ and $\|\nabla\widetilde{\eta}\|_{L^\infty(B_{(1+\delta)R})} \lesssim \frac{1}{\delta R}$. By Lemma 1 we can define the function $\widetilde{\eta}u \in H^1(\Omega)$ with the support property $\operatorname{supp}\widetilde{\eta}u \subset \overline{B_{(1+\delta)R} \cap \omega} = \overline{B}$ and therefore

$$\|\widetilde{\eta}u\|_{H^1(\Omega)} \le \|u\|_{L^2(B)} + \|\nabla(\widetilde{\eta}u)\|_{L^2(B)} \lesssim \frac{1}{\delta R}\|u\|_{L^2(B)} + \|\nabla u\|_{L^2(B)}. \qquad (27)$$

Then, we get

$$\begin{aligned}
\left|\left\langle \alpha u, \eta^2 u - J_h(\eta^2 u)\right\rangle_{L^2(\overline{B}\cap\Gamma_{\mathcal{R}})}\right| &= \left|\left\langle \alpha\widetilde{\eta}u, \eta^2 u - J_h(\eta^2 u)\right\rangle_{L^2(\overline{B}\cap\Gamma_{\mathcal{R}})}\right|\\
&\le \|\alpha\|_{L^\infty(\overline{B}\cap\Gamma_{\mathcal{R}})}\|\widetilde{\eta}u\|_{L^2(\overline{B}\cap\Gamma_{\mathcal{R}})}\left\|\eta^2 u - J_h(\eta^2 u)\right\|_{L^2(\overline{B}\cap\Gamma_{\mathcal{R}})}.
\end{aligned}$$

The multiplicative trace inequality (see, e.g., [9, inequality (1.6.2)]) for Ω and the estimate (27) gives

$$\|\widetilde{\eta}u\|_{L^2(\Gamma)} \lesssim \|\widetilde{\eta}u\|^{1/2}_{L^2(\Omega)}\|\widetilde{\eta}u\|^{1/2}_{H^1(\Omega)} \lesssim \frac{1}{\sqrt{\delta R}}\|u\|_{L^2(B)} + \|u\|^{1/2}_{L^2(B)}\|\nabla u\|^{1/2}_{L^2(B)}.$$

The multiplicative trace inequality for Ω and the estimates (25)–(26) imply

$$
\begin{aligned}
\|\eta^2 u - J_h(\eta^2 u)\|_{L^2(\Gamma)} &\lesssim \|\eta^2 u - J_h(\eta^2 u)\|_{L^2(\Omega)} + \|\eta^2 u - J_h(\eta^2 u)\|_{L^2(\Omega)}^{1/2} \|\nabla(\eta^2 u \\
&\quad - J_h(\eta^2 u))\|_{L^2(\Omega)}^{1/2} \lesssim \left(\frac{h^2}{\delta^2 R^2}\|u\|_{L^2(B)} + \frac{h^2}{\delta R}\|\nabla u\|_{L^2(B)}\right) \\
&\quad + \left(\frac{h}{\delta R}\|u\|_{L^2(B)}^{1/2} + \frac{h}{\sqrt{\delta R}}\|\nabla u\|_{L^2(B)}^{1/2}\right)\left(\frac{\sqrt{h}}{\delta R}\|u\|_{L^2(B)}^{1/2} + \frac{\sqrt{h}}{\sqrt{\delta R}}\|\nabla u\|_{L^2(B)}^{1/2}\right) \\
&\lesssim \frac{h^{3/2}}{(\delta R)^2}\|u\|_{L^2(B)} + \frac{h^{3/2}}{\delta R}\|\nabla u\|_{L^2(B)} + \frac{h^{3/2}}{(\delta R)^{3/2}}\|u\|_{L^2(B)}^{1/2}\|\nabla u\|_{L^2(B)}^{1/2} \\
&\lesssim \frac{h^{3/2}}{(\delta R)^2}\|u\|_{L^2(B)} + \frac{h^{3/2}}{\delta R}\|\nabla u\|_{L^2(B)}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\|\widetilde{\eta} u\|_{L^2(\Gamma)} \left\|\eta^2 u - J_h(\eta^2 u)\right\|_{L^2(\Gamma)} &\lesssim \left(\frac{1}{\sqrt{\delta R}}\|u\|_{L^2(B)} + \|u\|_{L^2(B)}^{1/2}\|\nabla u\|_{L^2(B)}^{1/2}\right) \\
&\quad \cdot \left(\frac{h^{3/2}}{(\delta R)^2}\|u\|_{L^2(B)} + \frac{h^{3/2}}{\delta R}\|\nabla u\|_{L^2(B)}\right) \\
&\lesssim \frac{h^{3/2}}{(\delta R)^{5/2}}\|u\|_{L^2(B)}^2 + \frac{h^{3/2}}{(\delta R)^{3/2}}\|u\|_{L^2(B)}\|\nabla u\|_{L^2(B)} \\
&\quad + \frac{h^{3/2}}{(\delta R)^2}\|u\|_{L^2(B)}^{3/2}\|\nabla u\|_{L^2(B)}^{1/2} + \frac{h^{3/2}}{\delta R}\|u\|_{L^2(B)}^{1/2}\|\nabla u\|_{L^2(B)}^{3/2}.
\end{aligned}
$$

Young's inequality and $h/(\delta R) \leq 1/4$, as well as $\delta \in (0, 1)$, $R \leq 2\,\mathrm{diam}(\Omega)$ implying $\frac{1}{\delta R} \lesssim \frac{1}{\delta^2 R^2}$, allow us to conclude (rather generously)

$$
\begin{aligned}
\left|\left\langle \alpha u, \eta^2 u - J_h(\eta^2 u)\right\rangle_{L^2(\overline{B}\cap\Gamma_{\mathcal{R}})}\right| &\lesssim \|\widetilde{\eta} u\|_{L^2(\Gamma)} \left\|\eta^2 u - J_h(\eta^2 u)\right\|_{L^2(\Gamma)} \\
&\lesssim \frac{h^2}{(\delta R)^2}\|\nabla u\|_{L^2(B)}^2 + \frac{1}{(\delta R)^2}\|u\|_{L^2(B)}^2 = \left(\frac{1+\delta}{\delta}\right)^2 |||u|||_{h,(1+\delta)R}^2. \qquad (28)
\end{aligned}
$$

Inserting the estimates (25), (26), (28) into (23) and with Young's inequality, we get with (22b) that

$$
\begin{aligned}
\|\nabla(\eta u)\|_{L^2(B)}^2 + \langle \alpha\eta u, \eta u\rangle_{L^2(\overline{B}\cap\Gamma_{\mathcal{R}})} &\lesssim a(u, \eta^2 u) + \frac{1}{\delta^2 R^2}\,\|u\|_{L^2(B)}^2 \\
&\lesssim \|\nabla u\|_{L^2(B)}\left(\frac{h}{\delta^2 R^2}\,\|u\|_{L^2(B)} + \frac{h}{\delta R}\,\|\eta\nabla u\|_{L^2(B)}\right) + \left(\|\nabla u\|_{L^2(B)} + \|\eta u\|_{L^2(B)}\right) \\
&\quad \times \left(\frac{h^2}{\delta^2 R^2}\,\|u\|_{L^2(B)} + \frac{h^2}{\delta R}\,\|\eta\nabla u\|_{L^2(B)}\right) + \frac{h^2}{\delta^2 R^2}\,\|\nabla u\|_{L^2(B)}^2 + \frac{1}{\delta^2 R^2}\,\|u\|_{L^2(B)}^2 \\
&\leq C\frac{h^2}{\delta^2 R^2}\,\|\nabla u\|_{L^2(B)}^2 + C\frac{1}{\delta^2 R^2}\,\|u\|_{L^2(B)}^2 + \frac{1}{2}\,\|\eta\nabla u\|_{L^2(B)}^2.
\end{aligned}
$$

Moving the term $\frac{1}{2}\|\eta\nabla u\|^2_{L^2(B)}$ to the left-hand side and inserting this estimate in (22a), we conclude the proof.

3.2 Low-dimensional approximation in $\mathcal{H}_h(D,\omega)$

In this subsection, we will derive a low dimensional approximation of the Galerkin solution by Scott–Zhang interpolation on a coarser grid.

We need to be able to extend functions defined on $B_{(1+2\delta)R}\cap\omega$ to $\mathbb{R}^d$. To this end, we use an extension operator $E : H^1(\Omega)\to H^1(\mathbb{R}^d)$, see, e.g., [1, Theorem 4.32], which satisfies $Eu = u$ on Ω and the H^1-stability estimate

$$\|Eu\|_{H^1(\mathbb{R}^d)} \leq C\,\|u\|_{H^1(\Omega)}\,.$$

For a function $u\in\mathcal{H}_h(B_{(1+2\delta)R},\omega)$ and a cut-off function $\eta\in C_0^\infty(B_{(1+2\delta)R})$ with $\operatorname{supp}\eta\subset B_{(1+\delta)R}$, $\eta\equiv 1$ on B_R we can define the function $\eta u\in H^1(\Omega)$ with the aid of Lemma 1. We note the support property $\operatorname{supp}(\eta u)\subset\overline{B_{(1+2\delta)R}\cap\omega}$, due to $\operatorname{supp}u\subset\overline{\omega}$. Therefore, the extension of ηu to Ω by zero is in $H^1(\Omega)$. Therefore, we have

$$\|E(\eta u)\|_{H^1(\mathbb{R}^d)} \leq C\,\|\eta u\|_{H^1(\omega)}\,. \tag{29}$$

Moreover, let $\Pi_{h,R} : (H^1(B_R\cap\omega), |||\cdot|||_{h,R}) \to (\mathcal{H}_h(B_R,\omega), |||\cdot|||_{h,R})$ be the orthogonal projection, which is well-defined since $\mathcal{H}_h(B_R,\omega)\subset H^1(B_R\cap\omega)$ is a closed subspace.

In the following lemma, we use a Scott–Zhang projection $I_H : H^1(\Omega)\to S^{1,1}(\mathcal{K}_H)$ of the form introduced in [41] for a quasiuniform grid $\mathcal{K}_H$ with mesh width H. By

$$\omega_K := \bigcup\left\{K'\in\mathcal{K}_H \;:\; K\cap K'\neq\emptyset\right\},$$

we denote the element patch of K, which contains K and all elements $K'\in\mathcal{K}_H$ that have a common node with K. Then, I_H has the following local approximation property for $u\in H^1(\omega_K)$

$$\|u - I_H u\|^2_{H^m(K)} \leq CH^{2(\ell-m)}\,|u|^2_{H^\ell(\omega_K)}\,,\ \ 0\leq\ell\leq 1. \tag{30}$$

The constant $C>0$ depends only on γ-shape regularity of the quasiuniform triangulation $\mathcal{K}_H$ and the dimension d.

Lemma 3 *Let $\delta\in(0,1)$, $R\in(0,2\operatorname{diam}(\Omega))$, B_R, $B_{(1+\delta)R}$, and $B_{(1+2\delta)R}$ be concentric boxes, and let $\omega\subseteq\Omega$ of the form (16) and $u\in\mathcal{H}_h(B_{(1+2\delta)R},\omega)$. Assume $\frac{h}{R}\leq\frac{\delta}{4}$. Let $\mathcal{K}_H$ be an (infinite) γ-shape regular triangulation of $\mathbb{R}^d$ and assume $\frac{H}{R}\leq\frac{\delta}{4}$ for the corresponding mesh width H. Let $\eta\in C_0^\infty(B_{(1+2\delta)R})$ be a cut-off function satisfying $\operatorname{supp}\eta\subset B_{(1+\delta)R}$, $\eta\equiv 1$ on B_R, and $\|\nabla\eta\|_{L^\infty(B_{(1+2\delta)R})}\lesssim\frac{1}{\delta R}$. Moreover, let $I_H : H^1(\mathbb{R}^d)\to S^{1,1}(\mathcal{K}_H)$ be the Scott–Zhang projection and $E : H^1(\Omega)\to H^1(\mathbb{R}^d)$ be an H^1-stable extension operator. Then, there exists a constant $C_{\text{app}}>0$, which depends only on the boundary value problem (3), Ω, d, p, the γ-shape regularity of the quasiuniform triangulation $\mathcal{T}_h$, and E such that*

(i) $(u - \Pi_{h,R} I_H E(\eta u))|_{B_R \cap \omega} \in \mathcal{H}_h(B_R, \omega)$;
(ii) $\left|\!\left|\!\left| u - \Pi_{h,R} I_H E(\eta u) \right|\!\right|\!\right|_{h,R} \le C_{\mathrm{app}} \frac{1+2\delta}{\delta} (\frac{h}{R} + \frac{H}{R}) \, |\!|\!| u |\!|\!|_{h,(1+2\delta)R}$;
(iii) $\dim W \le C_{\mathrm{app}} (\frac{(1+2\delta)R}{H})^d$, *where* $W := \Pi_{h,R} I_H E \mathcal{H}_h(B_{(1+2\delta)R}, \omega)$.

Proof The statement (iii) follows from the fact that $\dim I_H(E\mathcal{H}_h(B_{(1+2\delta)R}, \omega))|_{B_{(1+\delta)R}} \lesssim ((1+2\delta)R/H)^d$. For $u \in \mathcal{H}_h(B_{(1+2\delta)R}, \omega)$, we have $u|_{B_R \cap \omega} \in \mathcal{H}_h(B_R, \omega)$ as well and hence $\Pi_{h,R}(u|_{B_R \cap \omega}) = u|_{B_R \cap \omega}$, which gives (i). It remains to prove (ii): The assumption $\frac{H}{R} \le \frac{\delta}{4}$ implies $\bigcup \{K \in \mathcal{K}_H \ : \ \omega_K \cap B_R \neq \emptyset\} \subseteq B_{(1+\delta)R}$. The locality and the approximation properties (30) of I_H yield

$$\frac{1}{H} \left\| E(\eta u) - I_H E(\eta u) \right\|_{L^2(B_R)} + \left\| \nabla (E(\eta u) - I_H E(\eta u)) \right\|_{L^2(B_R)} \\ \lesssim \left\| \nabla E(\eta u) \right\|_{L^2(B_{(1+\delta)R})}.$$

We apply Lemma 2 with $\widetilde{R} = (1+\delta)R$ and $\widetilde{\delta} = \frac{\delta}{1+\delta}$. Note that $(1+\widetilde{\delta})\widetilde{R} = (1+2\delta)R$, and $\frac{h}{\widetilde{R}} \le \frac{\widetilde{\delta}}{4}$ follows from $4h \le \delta R = \widetilde{\delta}\widetilde{R}$. Hence, we obtain with (29)

$$\begin{aligned}
&\left|\!\left|\!\left| u - \Pi_{h,R} I_H E(\eta u) \right|\!\right|\!\right|_{h,R}^2 = \left|\!\left|\!\left| \Pi_{h,R} \left(E(\eta u) - I_H E(\eta u) \right) \right|\!\right|\!\right|_{h,R}^2 \le |\!|\!| E(\eta u) - I_H E(\eta u) |\!|\!|_{h,R}^2 \\
&= \left(\frac{h}{R}\right)^2 \left\| \nabla (E(\eta u) - I_H E(\eta u)) \right\|_{L^2(B_R \cap \omega)}^2 + \frac{1}{R^2} \left\| E(\eta u) - I_H E(\eta u) \right\|_{L^2(B_R \cap \omega)}^2 \\
&\lesssim \frac{h^2}{R^2} \left\| \nabla E(\eta u) \right\|_{L^2(B_{(1+\delta)R})}^2 + \frac{H^2}{R^2} \left\| \nabla E(\eta u) \right\|_{L^2(B_{(1+\delta)R})}^2 \lesssim \left(\frac{h^2}{R^2} + \frac{H^2}{R^2} \right) \left\| \eta u \right\|_{H^1(\Omega)}^2 \\
&\lesssim \left(\frac{h^2}{R^2} + \frac{H^2}{R^2} \right) \frac{1}{\delta^2 R^2} \left\| u \right\|_{L^2(B_{(1+\delta)R} \cap \omega)}^2 + \left(\frac{h^2}{R^2} + \frac{H^2}{R^2} \right) \left\| \nabla u \right\|_{L^2(B_{(1+\delta)R} \cap \omega)}^2 \\
&\lesssim \left(\frac{h^2}{R^2} + \frac{H^2}{R^2} \right) \frac{1}{\delta^2 R^2} \left\| u \right\|_{L^2(B_{(1+\delta)R} \cap \omega)}^2 + \left(\frac{h^2}{R^2} + \frac{H^2}{R^2} \right) \frac{(1+2\delta)^2}{\delta^2} |\!|\!| u |\!|\!|_{L^2(B_{(1+2\delta)R})}^2 \\
&\le \left(C_{\mathrm{app}} \frac{1+2\delta}{\delta} \left(\frac{h}{R} + \frac{H}{R} \right) \right)^2 |\!|\!| u |\!|\!|_{h,(1+2\delta)R}^2 ,
\end{aligned}$$

which concludes the proof.

By iterating this approximation result on suitable concentric boxes, we can derive a low-dimensional subspace in the space $\mathcal{H}_h$ and the bestapproximation in this space converges exponentially, which is stated in the following lemma.

Lemma 4 *Let C_{app} be the constant of Lemma 3. Let $q, \kappa \in (0, 1)$, $R \in (0, 2\operatorname{diam}(\Omega))$, $k \in \mathbb{N}$ and $\omega \subseteq \Omega$ be of the form (16). Assume*

$$\frac{h}{R} \le \frac{\kappa q}{8k \max\{1, C_{\mathrm{app}}\}}. \tag{31}$$

Then, there exists a subspace V_k of $S_0^{p,1}(\mathcal{T}_h, \Gamma_D)|_{B_R \cap \omega}$ with dimension

$$\dim V_k \le C_{\dim} \left(\frac{1+\kappa^{-1}}{q} \right)^d k^{d+1},$$

such that for every $u \in \mathcal{H}_h(B_{(1+\kappa)R}, \omega)$

$$\min_{v \in V_k} |||u - v|||_{h,R} \leq q^k \, |||u|||_{h,(1+\kappa)R} \,. \tag{32}$$

The constant $C_{\text{dim}} > 0$ depends only on the boundary value problem (3), Ω, d, and the γ-shape regularity of the quasiuniform triangulation $\mathcal{T}_h$.

Proof We iterate the approximation result of Lemma 3 on boxes $B_{(1+\delta_j)R}$, with $\delta_j := \kappa \frac{k-j}{k}$ for $j = 0, \ldots, k$. We note that $\kappa = \delta_0 > \delta_1 > \cdots > \delta_k = 0$. We choose $H = \frac{\kappa q R}{8k \max\{C_{\text{app}}, 1\}}$.

If $h \geq H$, then we select $V_k = \mathcal{H}_h(B_R, \omega)$. Due to the choice of H we have $\dim V_k \lesssim (\frac{R}{h})^d \lesssim k(\frac{R}{H})^d \simeq C_{\text{dim}}(\frac{1+\kappa^{-1}}{q})^d k^{d+1}$.

If $h < H$, we apply Lemma 3 with $\widetilde{R} = (1+\delta_j)R$ and $\widetilde{\delta}_j = \frac{1}{2k(1+\delta_j)} < \frac{1}{2}$. Note that $\delta_{j-1} = \delta_j + \frac{1}{k}$ gives $(1+\delta_{j-1})R = (1+2\widetilde{\delta}_j)\widetilde{R}$. The assumption $\frac{H}{\widetilde{R}} \leq \frac{1}{8k(1+\delta_j)} = \frac{\widetilde{\delta}_j}{4}$ is fulfilled due to our choice of H. For $j = 1$, Lemma 3 provides an approximation w_1 in a subspace W_1 of $\mathcal{H}_h(B_{(1+\delta_1)R}, \omega)$ with $\dim W_1 \leq C(\frac{(1+\kappa)R}{H})^d$ such that

$$\begin{aligned} |||u - w_1|||_{h,(1+\delta_1)R} &\leq 2C_{\text{app}} \frac{H}{(1+\delta_1)R} \frac{1+2\widetilde{\delta}_1}{\widetilde{\delta}_1} \, |||u|||_{h,(1+\delta_0)R} \\ &= 4C_{\text{app}} \frac{kH}{R}(1+2\widetilde{\delta}_1) \, |||u|||_{h,(1+\kappa)R} \leq q \, |||u|||_{h,(1+\kappa)R} \,. \end{aligned}$$

Since $u - w_1 \in \mathcal{H}_h(B_{(1+\delta_1)R}, \omega)$, we can use Lemma 3 again and get an approximation w_2 of $u - w_1$ in a subspace W_2 of $\mathcal{H}_h(B_{(1+\delta_2)R}, \omega)$ with $\dim W_2 \leq C(\frac{(1+\kappa)R}{H})^d$. Arguing as for $j = 1$, we get

$$|||u - w_1 - w_2|||_{h,(1+\delta_2)R} \leq q \, |||u - w_1|||_{h,(1+\delta_1)R} \leq q^2 \, |||u|||_{h,(1+\kappa)R} \,.$$

Continuing this process $k-2$ times leads to an approximation $v := \sum_{j=1}^k w_i$ in the space $V_k := \sum_{j=1}^k W_j$ of dimension $\dim V_k \leq Ck(\frac{(1+\kappa)R}{H})^d = C_{\text{dim}}(\frac{1+\kappa^{-1}}{q})^d k^{d+1}$.

Now we are able to prove the main result of this section.

Proof of Proposition 1 Choose $\kappa = \frac{1}{1+\eta}$. By assumption, we have $\operatorname{dist}(B_{R_\tau}, B_{R_\sigma}) \geq \eta^{-1} \operatorname{diam} B_{R_\tau} = \sqrt{d}\eta^{-1} R_\tau$. In particular, this implies

$$\begin{aligned} \operatorname{dist}(B_{(1+\kappa)R_\tau}, B_{R_\sigma}) &\geq \operatorname{dist}(B_{R_\tau}, B_{R_\sigma}) - \kappa R_\tau \sqrt{d} \geq \sqrt{d} R_\tau (\eta^{-1} - \kappa) \\ &= \sqrt{d} R_\tau \frac{1}{\eta(1+\eta)} > 0. \end{aligned}$$

The Galerkin solution ϕ_h satisfies $\phi_h|_{B_{(1+\delta)R} \cap \Omega} \in \mathcal{H}_h(B_{(1+\delta)R}, \Omega)$. The coercivity (5) of the bilinear form $a(\cdot,\cdot)$ implies

$$\|\phi_h\|^2_{H^1(\Omega)} \lesssim a(\phi_h, \phi_h) = \langle f, \phi_h \rangle = \left\langle \Pi^{L^2} f, \phi_h \right\rangle \lesssim \left\| \Pi^{L^2} f \right\|_{L^2(\Omega)} \|\phi_h\|_{H^1(\Omega)} \,.$$

Furthermore, with $\frac{h}{R_\tau} < 1$, we get

$$\|\!|\phi_h|\!\|_{h,(1+\kappa)R_\tau} \lesssim \left(1+\frac{1}{R_\tau}\right)\|\phi_h\|_{H^1(\Omega)} \lesssim \left(1+\frac{1}{R_\tau}\right)\left\|\Pi^{L^2} f\right\|_{L^2(\Omega)},$$

and we have a bound on the right-hand side of (32). We are now in the position to define the space V_k, for which we distinguish two cases.

Case 1: The condition (31) is satisfied with $R = R_\tau$. With the space V_k provided by Lemma 4 we get

$$\begin{aligned}\min_{v\in V_k}\|\phi_h - v\|_{L^2(B_{R_\tau}\cap\Omega)} \leq R_\tau \min_{v\in V_k}\|\!|\phi_h - v|\!\|_{h,R_\tau} &\lesssim (R_\tau+1)q^k\left\|\Pi^{L^2} f\right\|_{L^2(\Omega)}\\ &\lesssim \operatorname{diam}(\Omega)q^k\left\|\Pi^{L^2} f\right\|_{L^2(\Omega)},\end{aligned}$$

and the dimension of V_k is bounded by $\dim V_k \leq C((2+\eta)q^{-1})^d k^{d+1}$.

Case 2: The condition (31) is not satisfied, i.e., we have $\frac{h}{R_\tau} \geq \frac{\kappa q}{8k\max\{1,C_{\mathrm{app}}\}}$.

Then, we select $V_k := \{v|_{B_{R_\tau}\cap\Omega} : v \in S_0^{p,1}(\mathcal{T}_h,\Gamma_D)\}$, and the minimum in (14) is obviously zero. By choice of κ, the dimension of V_k is bounded by

$$\dim V_k \lesssim \left(\frac{R_\tau}{h}\right)^d \lesssim \left(\frac{8k\max\{C_{\mathrm{app}},1\}}{\kappa q}\right)^d \lesssim \left((1+\eta)q^{-1}\right)^d k^{d+1},$$

which concludes the proof of the non trivial statement in (14). The other estimate follows directly from the $L^2(\Omega)$-stability of the $L^2(\Omega)$-orthogonal projection.

Remark 4 A result similar to Proposition 1 holds for the pure Neumann problem, i.e., $\Gamma = \Gamma_{\mathcal{N}}$ as well. In this case, the matrix $\mathbf{A}$ is not invertible and therefore either a stabilized Galerkin discretization or a saddle point formulation has to be chosen to deal with the one-dimensional kernel. The general ideas underlying Proposition 1 can be utilized, see [15,17] for details.

4 Proof of main results

We use the approximation of ϕ_h from the low dimensional spaces given in Proposition 1 to construct a blockwise low-rank approximation of $\mathbf{A}^{-1}$ and in turn an $\mathcal{H}$-matrix approximation of $\mathbf{A}^{-1}$. In fact, we will only use a FEM-isomorphism to transfer Proposition 1 to the matrix level, which follows the lines of [6, Theorem 2].

Proof of Theorem 1 If $C_{\dim}(2+\eta)^d q^{-d}k^{d+1} \geq \min(|\tau|,|\sigma|)$, we use the exact matrix block $\mathbf{X}_{\tau\sigma} = \mathbf{A}^{-1}|_{\tau\times\sigma}$ and $\mathbf{Y}_{\tau\sigma} = \mathbf{I} \in \mathbb{R}^{|\sigma|\times|\sigma|}$.

If $C_{\dim}(2+\eta)^d q^{-d}k^{d+1} < \min(|\tau|,|\sigma|)$, let $\lambda_i : L^2(\Omega) \to \mathbb{R}$ be continuous linear functionals on $L^2(\Omega)$ satisfying $\lambda_i(\psi_j) = \delta_{ij}$. We define $\mathbb{R}^\tau := \{\mathbf{x} \in \mathbb{R}^N : x_i = 0\ \forall\, i \notin \tau\}$ and the mappings

$$\Lambda_\tau : L^2(\Omega) \to \mathbb{R}^\tau, v \mapsto (\lambda_i(v))_{i\in\tau} \text{ and } \mathcal{J}_\tau : \mathbb{R}^\tau \to S_0^{p,1}(\mathcal{T}_h, \Gamma_D),\ \mathbf{x} \mapsto \sum_{j\in\tau} x_j \psi_j.$$

For $\mathbf{x} \in \mathbb{R}^\tau$, (6) leads to the stability estimate

$$h^{d/2}\|\mathbf{x}\|_2 \lesssim \|\mathcal{J}_\tau \mathbf{x}\|_{L^2(\Omega)} \lesssim h^{d/2}\|\mathbf{x}\|_2 . \tag{33}$$

Let V_k be the finite dimensional subspace from Proposition 1.

Because of (33) and the L^2-stability of $\mathcal{J}_\mathcal{I}\Lambda_\mathcal{I}$, the adjoint $\Lambda_\mathcal{I}^* : \mathbb{R}^N \to L^2(\Omega)' \simeq L^2(\Omega)$ of $\Lambda_\mathcal{I}$ satisfies

$$\begin{aligned}\left\|\Lambda_\mathcal{I}^*\mathbf{b}\right\|_{L^2(\Omega)} &= \sup_{v\in L^2(\Omega)} \frac{\langle \mathbf{b}, \Lambda_\mathcal{I} v\rangle_2}{\|v\|_{L^2(\Omega)}} \lesssim \|\mathbf{b}\|_2 \sup_{v\in L^2(\Omega)} \frac{h^{-d/2}\|\mathcal{J}_\mathcal{I}\Lambda_\mathcal{I} v\|_{L^2(\Omega)}}{\|v\|_{L^2(\Omega)}} \\ &\le Ch^{-d/2}\|\mathbf{b}\|_2 .\end{aligned}$$

Moreover, if $\mathbf{b} = (\langle f, \psi_i\rangle)_{i\in\mathcal{I}}$, we have $(\Lambda_\mathcal{I}^*\mathbf{b})(\psi_i) = b_i = \langle f, \psi_i\rangle = \langle \Pi^{L^2} f, \psi_i\rangle$. Therefore, f and $\Lambda_\mathcal{I}^*\mathbf{b} = \Pi^{L^2} f$ have the same Galerkin approximation.

Let V_k be the finite dimensional subspace from Proposition 1. We define $\mathbf{X}_{\tau\sigma}$ as an orthogonal basis of the space $\mathcal{V}_\tau := \{\Lambda_\tau v : v \in V_k\}$. Then, the rank of $\mathbf{X}_{\tau\sigma}$ is bounded by $\dim V_k \le C_{\dim}(2+\eta)^d q^{-d} k^{d+1}$.

The estimate (33) and the approximation result from Proposition 1 provide the error estimate

$$\begin{aligned}\|\Lambda_\tau\phi_h - \Lambda_\tau v\|_2 &\lesssim h^{-d/2}\|\mathcal{J}_\tau(\Lambda_\tau\phi_h - \Lambda_\tau v)\|_{L^2(\Omega)} \lesssim h^{-d/2}\|\phi_h - v\|_{L^2(B_{R_\tau}\cap\Omega)} \\ &\le C_{\text{box}} h^{-d/2} q^k \left\|\Pi^{L^2} f\right\|_{L^2(\Omega)} \lesssim C_{\text{box}} h^{-d} q^k \|\mathbf{b}\|_2 .\end{aligned}$$

Since $\mathbf{X}_{\tau\sigma}\mathbf{X}_{\tau\sigma}^T$ is the orthogonal projection from $\mathbb{R}^N$ onto $\mathcal{V}_\tau$, we get that $z := \mathbf{X}_{\tau\sigma}\mathbf{X}_{\tau\sigma}^T\Lambda_\tau\phi_h$ is the best approximation of $\Lambda_\tau\phi_h$ in $\mathcal{V}_\tau$ and arrive at

$$\|\Lambda_\tau\phi_h - z\|_2 \le \|\Lambda_\tau\phi_h - \Lambda_\tau v\|_2 \lesssim C_{\text{box}} N q^k \|\mathbf{b}\|_2 .$$

If we define $\mathbf{Y}_{\tau,\sigma} := \mathbf{A}^{-1}|_{\tau\times\sigma}^T \mathbf{X}_{\tau\sigma}$, we get $z = \mathbf{X}_{\tau\sigma}\mathbf{Y}_{\tau\sigma}^T\mathbf{b}$, since $\Lambda_\tau\phi_h = \mathbf{A}^{-1}|_{\tau\times\sigma}\mathbf{b}$.

The following lemma gives an estimate for the global spectral norm by the local spectral norms, which we will use in combination with Theorem 1 to derive our main result, Theorem 2.

Lemma 5 [7,20,28, Lemma 6.5.8] *Let* $\mathbf{M} \in \mathbb{R}^{N\times N}$ *and* P *be a partitioning of* $\mathcal{I}\times\mathcal{I}$. *Then,*

$$\|\mathbf{M}\|_2 \le C_{\text{sp}}\left(\sum_{\ell=0}^{\infty} \max\{\|\mathbf{M}|_{\tau\times\sigma}\|_2 : (\tau,\sigma) \in P, \mathit{level}(\tau) = \ell\}\right).$$

Now we are able to prove our main result, Theorem 2.

Proof of Theorem 2 For each admissible cluster pair (τ, σ), Theorem 1 provides matrices $\mathbf{X}_{\tau\sigma} \in \mathbb{R}^{|\tau|\times r}$, $\mathbf{Y}_{\tau\sigma} \in \mathbb{R}^{r\times|\sigma|}$, so that we can define the $\mathcal{H}$-matrix $\mathbf{V}_{\mathcal{H}}$ by

$$\mathbf{B}_{\mathcal{H}} = \begin{cases} \mathbf{X}_{\tau\sigma}\mathbf{Y}_{\tau\sigma}^T & \text{if } (\tau,\sigma) \in P_{\text{far}}, \\ \mathbf{A}^{-1}|_{\tau\times\sigma} & \text{otherwise.} \end{cases}$$

On each admissible block $(\tau, \sigma) \in P_{\text{far}}$, we can use the blockwise estimate of Theorem 1 and get for $q \in (0, 1)$

$$\left\|(\mathbf{A}^{-1} - \mathbf{B}_{\mathcal{H}})|_{\tau\times\sigma}\right\|_2 \leq C_{\text{apx}} N q^k.$$

On inadmissible blocks, the error is zero by definition. Therefore, Lemma 5 concludes the proof, since

$$\begin{aligned}\left\|\mathbf{A}^{-1} - \mathbf{B}_{\mathcal{H}}\right\|_2 &\leq C_{\text{sp}}\left(\sum_{\ell=0}^{\infty} \max\left\{\left\|(\mathbf{A}^{-1} - \mathbf{B}_{\mathcal{H}})|_{\tau\times\sigma}\right\|_2 : (\tau,\sigma) \in P, \text{level}(\tau) = \ell\right\}\right) \\ &\leq C_{\text{apx}} C_{\text{sp}} N q^k \text{depth}(\mathbb{T}_{\mathcal{I}}).\end{aligned}$$

Since in Theorem 1 we have $r \leq C_{\text{dim}}(2+\eta)^d q^{-d} k^{d+1}$, defining $b = -\frac{\ln(q)}{C_{\text{dim}}^{1/(d+1)}} q^{d/(d+1)}(2+\eta)^{-d/(1+d)} > 0$, we obtain $q^k = e^{-br^{1/(d+1)}}$ and hence

$$\left\|\mathbf{A}^{-1} - \mathbf{B}_{\mathcal{H}}\right\|_2 \leq C_{\text{apx}} C_{\text{sp}} N \text{depth}(\mathbb{T}_{\mathcal{I}}) e^{-br^{1/(d+1)}},$$

which concludes the proof.

5 Hierarchical LU-decomposition

In [3] the existence of an (approximate) $\mathcal{H}$-LU decomposition, i.e., a factorization of the form $\mathbf{A} \approx \mathbf{L}_{\mathcal{H}}\mathbf{U}_{\mathcal{H}}$ with lower and upper triangular $\mathcal{H}$-matrices $\mathbf{L}_{\mathcal{H}}$ and $\mathbf{U}_{\mathcal{H}}$, was asserted for finite element matrices $\mathbf{A}$ corresponding to the Dirichlet problem for elliptic operators with L^∞-coefficients. In [25] this result was extended to the case, where the block structure of the $\mathcal{H}$-matrix is constructed by domain decomposition clustering methods, instead of the standard geometric bisection clustering.

Algorithms for computing an $\mathcal{H}$-LU decomposition have been proposed repeatedly in the literature, e.g., [2,37] and numerical evidence for their usefulness put forward; we mention here that $\mathcal{H}$-LU decomposition can be employed for black box preconditioning in iterative solvers, [2,21,23,24,35]. An existence result for $\mathcal{H}$-LU factorization is then an important step towards a mathematical understanding of the good performance of these schemes.

The main steps in the proof of [3] are to approximate certain Schur complements of $\mathbf{A}$ by $\mathcal{H}$-matrices and to show a recursion formula for the Schur complement. Using these two observations an approximation of the exact LU-factors for the Schur complements, and consequently for the whole matrix, can be derived recursively.

Since the construction of the approximate LU-factors is completely algebraic, once we know that the Schur complements have an $\mathcal{H}$-matrix approximation of arbitrary accuracy, we will show that we can provide such an approximation and only sketch the remaining steps. Details can be found in [3,25].

Our main result, Theorem 2, shows the existence of an $\mathcal{H}$-matrix approximation to the inverse FEM stiffness matrix with arbitrary accuracy, whereas previous results achieve accuracy up to the finite element error. In fact, both [3,25] assume, in order to derive an $\mathcal{H}$-LU decomposition, that approximations to the inverse with arbitrary accuracy exist. Thus, due to our main result this assumption is fulfilled for inverse finite element matrices for elliptic operators with various boundary conditions.

Since we are in the setting of the Lax–Milgram Lemma, we get that the, in general, non symmetric matrix $\mathbf{A}$ is positive definite in the sense that $\mathbf{x}^T\mathbf{A}\mathbf{x} > 0$ for all $\mathbf{x} \neq 0$. Therefore, $\mathbf{A}$ has an LU-decomposition $\mathbf{A} = \mathbf{L}\mathbf{U}$, where $\mathbf{L}$ is a lower triangular matrix and $\mathbf{U}$ is an upper triangular matrix, independently of the numbering of the degrees of freedom, i.e., every other numbering of the basis functions permits an LU-decomposition as well (see, e.g., [33, Cor. 3.5.6]). By classical linear algebra (see, e.g., [33, Cor. 3.5.6]), this implies that for any $n \leq N$ and index set $\rho := \{1, \dots, n\}$, the matrix $\mathbf{A}|_{\rho\times\rho}$ is invertible.

We start with the approximation of appropriate Schur complements.

5.1 Schur complements

One way to approximate the Schur complement for a finite element matrix is to follow the lines of [3,25] by using $\mathcal{H}$-arithmetics and the sparsity of the finite element matrix. We present a different way of deriving such a result, which is more in line with our procedure in Sect. 3. It relies on interpreting Schur complements as FEM stiffness matrices from constrained spaces.

Lemma 6 *Let (τ, σ) be an admissible cluster pair and $\rho := \{i \in \mathcal{I} : i < \min(\tau\cup\sigma)\}$. Define the Schur complement $\mathbf{S}(\tau, \sigma) = \mathbf{A}|_{\tau\times\sigma} - \mathbf{A}|_{\tau\times\rho}(\mathbf{A}|_{\rho\times\rho})^{-1}\mathbf{A}|_{\rho\times\sigma}$. Then, there exists a rank-r matrix $\mathbf{S}_{\mathcal{H}}(\tau, \sigma)$ such that*

$$\|\mathbf{S}(\tau, \sigma) - \mathbf{S}_{\mathcal{H}}(\tau, \sigma)\|_2 \leq C_{\text{sc}} h^{-1} e^{-br^{1/(d+1)}} \|\mathbf{A}\|_2\,,$$

where the constant $C_{\text{sc}} > 0$ depends only on the boundary value problem (3), Ω, p, d, and the γ-shape regularity of the quasiuniform triangulation $\mathcal{T}_h$.

Proof We define $\omega_\rho = \text{interior}(\bigcup_{i\in\rho} \operatorname{supp} \psi_i) \subset \Omega$ and let B_{R_τ}, B_{R_σ} be bounding boxes for the clusters τ, σ with (8). Our starting point is the well-known observation that the Schur complement matrix $\mathbf{S}(\tau, \sigma)$ can be understood in terms of an orthogonalization with respect to the degrees of freedom in ρ. That is, for $\mathbf{u} \in \mathbb{R}^{|\tau|}$, $\mathbf{w} \in \mathbb{R}^{|\sigma|}$ a direct calculation (see, e.g., [10] for the essentials) shows

$$\mathbf{u}^T\mathbf{S}(\tau, \sigma)\mathbf{w} = a(\widetilde{u}, w), \tag{34}$$

with $w = \sum_{j=1}^{|\sigma|} \mathbf{w}_j \psi_{j_\sigma}$, where the index j_σ denotes the j-th basis function corresponding to the cluster σ, and the function $\widetilde{u} \in S_0^{p,1}(\mathcal{T}_h, \Gamma_D)$ is defined by $\widetilde{u} = \sum_{j=1}^{|\tau|} \mathbf{u}_j \psi_{j_\tau} + u_\rho$ with $\operatorname{supp} u_\rho \subset \overline{\omega_\rho}$ such that

$$a(\widetilde{u}, w) = 0 \quad \forall w \in S_0^{p,1}(\mathcal{T}_h, \Gamma_D) \text{ with } \operatorname{supp} w \subset \overline{\omega_\rho}. \tag{35}$$

The key to approximate the Schur complement $\mathbf{S}(\tau, \sigma)$ is to approximate the function $\widetilde{u}$. We will provide such an approximation by applying the techniques from the previous sections with the use of the orthogonality (35).

Since $\operatorname{supp} \widetilde{u} \subset B_{R_\tau} \cup \overline{\omega_\rho}$, we get for w with $\operatorname{supp} w \subset B_{R_\sigma}$ that

$$a(\widetilde{u}, w) = a(\widetilde{u}|_{\operatorname{supp} w}, w) = a(\widetilde{u}|_{B_{R_\sigma} \cap \omega_\rho}, w).$$

Therefore, we only need to approximate $\widetilde{u}$ on the intersection $B_{R_\sigma} \cap \omega_\rho$. This support property and the orthogonality (35) imply that $\widetilde{u} \in \mathcal{H}_h(B_{(1+\delta)R_\sigma}, \omega_\rho)$.

Therefore, Lemma 2 can be applied to $\widetilde{u}$. As a consequence, Lemma 4 provides a low dimensional space V_k, where the choice $\kappa = \frac{1}{\eta+1}$ bounds the dimension of V_k by $\dim V_k \leq C_{\dim}(2+\eta)^d q^{-d} k^{d+1}$. Moreover, the best approximation $\widetilde{v} = \Pi_{V_k} \widetilde{u} \in V_k$ to $\widetilde{u}$ in the space V_k satisfies

$$|\!|\!| \widetilde{u} - \widetilde{v} |\!|\!|_{h,(1+\delta)R_\sigma} \leq q^k \, |\!|\!| \widetilde{u} |\!|\!|_{h,(1+\delta)R_\sigma} .$$

This implies

$$\begin{aligned} |a(\widetilde{u}, w) - a(\widetilde{v}, w)| &\lesssim \|\widetilde{u} - \widetilde{v}\|_{H^1(B_{(1+\delta)R_\sigma} \cap \omega_\rho)} \|w\|_{H^1(B_{(1+\delta)R_\sigma} \cap \Omega)} \\ &\lesssim \frac{R_\sigma}{h} |\!|\!| \widetilde{u} - \widetilde{v} |\!|\!|_{h,(1+\delta)R_\sigma} \|w\|_{H^1(\Omega)} \lesssim h^{-1} q^k \|\widetilde{u}\|_{H^1(\Omega)} \|w\|_{H^1(\Omega)} . \end{aligned}$$

Since $\operatorname{supp}(\widetilde{u} - u) = \operatorname{supp}(u_\rho) \subset \overline{\omega_\rho}$ with $u = \sum_{j=1}^{|\tau|} \mathbf{u}_j \psi_{j_\tau}$, the coercivity (5) and orthogonality (35) lead to

$$\|\widetilde{u} - u\|_{H^1(\Omega)}^2 \lesssim a(\widetilde{u} - u, \widetilde{u} - u) = a(-u, \widetilde{u} - u) \lesssim \|u\|_{H^1(\Omega)} \|\widetilde{u} - u\|_{H^1(\Omega)} .$$

Consequently, we get with an inverse estimate and (33) that

$$\begin{aligned} |a(\widetilde{u}, w) - a(\widetilde{v}, w)| &\lesssim h^{-1} q^k \left(\|\widetilde{u} - u\|_{H^1(\Omega)} + \|u\|_{H^1(\Omega)} \right) \|w\|_{H^1(\Omega)} \\ &\lesssim h^{-1} q^k \|u\|_{H^1(\Omega)} \|w\|_{H^1(\Omega)} \lesssim h^{d-3} q^k \|\mathbf{u}\|_2 \|\mathbf{w}\|_2 . \end{aligned}$$

The linear mapping $\mathcal{E} : u \mapsto \widetilde{v}$ with $\dim \operatorname{ran} \mathcal{E} \leq C_{\dim}(2+\eta)^d q^{-d} k^{d+1}$ has a matrix representation $\mathbf{u} \mapsto \mathbf{B}\mathbf{u}$, where the rank of $\mathbf{B}$ is bounded by $C_{\dim}(2+\eta)^d q^{-d} k^{d+1}$. Therefore, we get that $a(\mathcal{E}u, w) = \mathbf{u}^T \mathbf{B}^T \mathbf{A}|_{\tau \times \sigma} \mathbf{w}$. The definition $\mathbf{S}_{\mathcal{H}}(\tau, \sigma) := \mathbf{B}^T \mathbf{A}|_{\tau \times \sigma}$ leads to a matrix $\mathbf{S}_{\mathcal{H}}(\tau, \sigma)$ of rank $r \leq C_{\dim}(2+\eta)^d q^{-d} k^{d+1}$ such that

$$\begin{aligned}\|\mathbf{S}(\tau,\sigma)-\mathbf{S}_{\mathcal{H}}(\tau,\sigma)\|_2 &= \sup_{\mathbf{u}\in\mathbb{R}^{|\tau|},\mathbf{w}\in\mathbb{R}^{|\sigma|}} \frac{\left|\mathbf{u}^T(\mathbf{S}(\tau,\sigma)-\mathbf{S}_{\mathcal{H}}(\tau,\sigma))\mathbf{w}\right|}{\|\mathbf{u}\|_2\,\|\mathbf{w}\|_2} \\ &\leq Ch^{d-3}e^{-br^{1/(d+1)}},\end{aligned}$$

and the estimate $\frac{1}{\|\mathbf{A}\|_2}\lesssim h^{2-d}$ from [14, Theorem 2] finishes the proof.

We refer to the next subsection for the existence of the inverse $\mathbf{S}(\tau,\tau)^{-1}$ of the Schur complement $\mathbf{S}(\tau,\tau)$. We proceed to approximate it by blockwise rank-r matrices. With the representation of the Schur complement from (34), we get that for a given right-hand side $f\in L^2(\Omega)$, solving $\mathbf{S}(\tau,\tau)\mathbf{u}=\mathbf{f}$ with $\mathbf{f}\in\mathbb{R}^{|\tau|}$ defined by $\mathbf{f}_i=\left\langle f,\psi_{i_\tau}\right\rangle$, is equivalent to solving $a(\widetilde{u},w)=\langle f,w\rangle$ for all $w\in S_0^{p,1}(\mathcal{T}_h,\Gamma_D)$ with supp $w\subset\overline{\omega_\tau}$. Let $\tau_1\times\sigma_1\subset\tau\times\tau$ be an η-admissible subblock. For $f\in L^2(\Omega)$ with supp $f\subset B_{R_{\sigma_1}}$, we get the orthogonality

$$a(\widetilde{u},w)=0\quad\forall w\in S_0^{p,1}(\mathcal{T}_h,\Gamma_D),\,\text{supp}\,w\subset B_{R_{\tau_1}}\cap\overline{\omega_\tau}.$$

Therefore, we have $\widetilde{u}\in\mathcal{H}_h(B_{R_{\tau_1}},\omega_\tau)$ and our results from Sect. 3 can be applied to approximate $\widetilde{u}$ on $B_{R_{\tau_1}}\cap\omega_\tau$. As in Sect. 4, this approximation can be used to construct a rank-r factorization of the subblock $\mathbf{S}(\tau,\tau)^{-1}|_{\tau_1\times\sigma_1}$, which is stated in the following theorem.

Theorem 3 *Let $\tau\subset\mathcal{I}$ and $\rho:=\{i\in\mathcal{I}:i<\min(\tau)\}$ and $\tau_1\times\sigma_1\subset\tau\times\tau$ be η-admissible. Define the Schur complement $\mathbf{S}(\tau,\tau)=\mathbf{A}|_{\tau\times\tau}-\mathbf{A}|_{\tau\times\rho}(\mathbf{A}|_{\rho\times\rho})^{-1}\mathbf{A}|_{\rho\times\tau}$. Then, there exist rank-r matrices $\mathbf{X}_{\tau_1\sigma_1}\in\mathbb{R}^{|\tau_1|\times r}$, $\mathbf{Y}_{\tau_1\sigma_1}\in\mathbb{R}^{|\sigma_1|\times r}$ such that*

$$\left\|\mathbf{S}(\tau,\tau)^{-1}|_{\tau_1\times\sigma_1}-\mathbf{X}_{\tau_1\sigma_1}\mathbf{Y}_{\tau_1\sigma_1}^T\right\|_2\leq C_{\text{apx}}Ne^{-br^{1/(d+1)}}. \tag{36}$$

The constants $C_{\text{apx}}, b>0$ depend only on the boundary value problem (3), Ω, d, p, and the γ-shape regularity of the quasiuniform triangulation $\mathcal{T}_h$.

5.2 Existence of $\mathcal{H}$-LU decomposition

In this subsection, we will use the approximation of the Schur complement from the previous section to prove the existence of an (approximate) $\mathcal{H}$-LU decomposition. We start with a hierarchical relation of the Schur complements $\mathbf{S}(\tau,\tau)$.

The Schur complements $\mathbf{S}(\tau,\tau)$ for a block $\tau\in\mathbb{T}_{\mathcal{I}}$ can be derived from the Schur complements of its sons by

$$\mathbf{S}(\tau,\tau)=\begin{pmatrix}\mathbf{S}(\tau_1,\tau_1) & \mathbf{S}(\tau_1,\tau_2)\\ \mathbf{S}(\tau_2,\tau_1) & \mathbf{S}(\tau_2,\tau_2)+\mathbf{S}(\tau_2,\tau_1)\mathbf{S}(\tau_1,\tau_1)^{-1}\mathbf{S}(\tau_1,\tau_2)\end{pmatrix},$$

where τ_1,τ_2 are the sons of τ. A proof of this relation can be found in [3, Lemma 3.1]. One should note that the proof does not use any properties of the matrix $\mathbf{A}$ other than invertibility and existence of an LU-decomposition. Moreover, we have by definition of $\mathbf{S}(\tau,\tau)$ that $\mathbf{S}(\mathcal{I},\mathcal{I}) = \mathbf{A}$.

If τ is a leaf, we get the LU-decomposition of $\mathbf{S}(\tau,\tau)$ by the classical LU-decomposition, which exists since $\mathbf{A}$ has an LU-decomposition. If τ is not a leaf, we use the hierarchical relation of the Schur complements to define an LU-decomposition of the Schur complement $\mathbf{S}(\tau,\tau)$ by

$$\mathbf{L}(\tau) := \begin{pmatrix} \mathbf{L}(\tau_1) & 0 \\ \mathbf{S}(\tau_2,\tau_1)\mathbf{U}(\tau_1)^{-1} & \mathbf{L}(\tau_2) \end{pmatrix}, \quad \mathbf{U}(\tau) := \begin{pmatrix} \mathbf{U}(\tau_1) & \mathbf{L}(\tau_1)^{-1}\mathbf{S}(\tau_1,\tau_2) \\ 0 & \mathbf{U}(\tau_2) \end{pmatrix}, \tag{37}$$

with $\mathbf{S}(\tau_1,\tau_1) = \mathbf{L}(\tau_1)\mathbf{U}(\tau_1)$, $\mathbf{S}(\tau_2,\tau_2) = \mathbf{L}(\tau_2)\mathbf{U}(\tau_2)$ and indeed get $\mathbf{S}(\tau,\tau) = \mathbf{L}(\tau)\mathbf{U}(\tau)$. Moreover, the uniqueness of the LU-decomposition of $\mathbf{A}$ implies that due to $\mathbf{LU} = \mathbf{A} = \mathbf{S}(\mathcal{I},\mathcal{I}) = \mathbf{L}(\mathcal{I})\mathbf{U}(\mathcal{I})$, we have $\mathbf{L} = \mathbf{L}(\mathcal{I})$ and $\mathbf{U} = \mathbf{U}(\mathcal{I})$.

The existence of the inverses $\mathbf{L}(\tau_1)^{-1}$ and $\mathbf{U}(\tau_1)^{-1}$ follows by induction over the levels, since on a leaf the existence is clear and the matrices $\mathbf{L}(\tau)$, $\mathbf{U}(\tau)$ are block triangular matrices. Consequently, the inverse of $\mathbf{S}(\tau,\tau)$ exists.

Moreover, the restriction of the lower triangular part $\mathbf{S}(\tau_2,\tau_1)\mathbf{U}(\tau_1)^{-1}$ of the matrix $\mathbf{L}(\tau)$ to a subblock $\tau_2' \times \tau_1'$ with τ_i' a son of τ_i satisfies

$$\left(\mathbf{S}(\tau_2,\tau_1)\mathbf{U}(\tau_1)^{-1}\right)|_{\tau_2'\times\tau_1'} = \mathbf{S}(\tau_2',\tau_1')\mathbf{U}(\tau_1')^{-1},$$

and the upper triangular part of $\mathbf{U}(\tau)$ satisfies a similar relation.

The following Lemma shows that the spectral norm of the inverses $\mathbf{L}(\tau)^{-1}$, $\mathbf{U}(\tau)^{-1}$ can be bounded by the norm of the inverses $\mathbf{L}(\mathcal{I})^{-1}$, $\mathbf{U}(\mathcal{I})^{-1}$.

Lemma 7 *For $\tau \in \mathbb{T}_{\mathcal{I}}$, let $\mathbf{L}(\tau)$, $\mathbf{U}(\tau)$ be given by (37). Then,*

$$\max_{\tau\in\mathbb{T}_{\mathcal{I}}} \left\|\mathbf{L}(\tau)^{-1}\right\|_2 = \left\|\mathbf{L}(\mathcal{I})^{-1}\right\|_2,$$
$$\max_{\tau\in\mathbb{T}_{\mathcal{I}}} \left\|\mathbf{U}(\tau)^{-1}\right\|_2 = \left\|\mathbf{U}(\mathcal{I})^{-1}\right\|_2.$$

Proof We only show the result for $\mathbf{L}(\tau)$. With the block structure of (37) we get the inverse

$$\mathbf{L}(\tau)^{-1} = \begin{pmatrix} \mathbf{L}(\tau_1)^{-1} & 0 \\ -\mathbf{L}(\tau_2)^{-1}\mathbf{S}(\tau_2,\tau_1)\mathbf{U}(\tau_1)^{-1}\mathbf{L}(\tau_1)^{-1} & \mathbf{L}(\tau_2)^{-1} \end{pmatrix}.$$

So, we get by choosing $\mathbf{x}$ such that $\mathbf{x}_i = 0$ for $i \in \tau_1$ that

$$\left\|\mathbf{L}(\tau)^{-1}\right\|_2 = \sup_{\mathbf{x}\in\mathbb{R}^{|\tau|},\|x\|_2=1} \left\|\mathbf{L}(\tau)^{-1}\mathbf{x}\right\|_2 \geq \sup_{\mathbf{x}\in\mathbb{R}^{|\tau_2|},\|x\|_2=1} \left\|\mathbf{L}(\tau_2)^{-1}\mathbf{x}\right\|_2 = \left\|\mathbf{L}(\tau_2)^{-1}\right\|_2.$$

The same argument for $(\mathbf{L}(\tau)^{-1})^T$ leads to

$$\left\|\mathbf{L}(\tau)^{-1}\right\|_2 = \left\|\left(\mathbf{L}(\tau)^{-1}\right)^T\right\|_2 \geq \left\|\mathbf{L}(\tau_1)^{-1}\right\|_2.$$

Thus, we have $\left\|\mathbf{L}(\tau)^{-1}\right\|_2 \geq \max_{i=1,2}\left\|\mathbf{L}(\tau_1)^{-1}\right\|_2$ and as a consequence $\max_{\tau\in\mathbb{T}_\mathcal{I}}\left\|\mathbf{L}(\tau)^{-1}\right\|_2 = \left\|\mathbf{L}(\mathcal{I})^{-1}\right\|_2$.

We can now formulate the existence result for an $\mathcal{H}$-LU decomposition.

Theorem 4 *Let* $\mathbf{A} = \mathbf{LU}$ *with* $\mathbf{L}, \mathbf{U}$ *being lower and upper triangular matrices. There exist lower and upper triangular blockwise rank-*r *matrices* $\mathbf{L}_\mathcal{H}, \mathbf{U}_\mathcal{H}$ *such that*

$$\begin{aligned}\|\mathbf{A}-\mathbf{L}_\mathcal{H}\mathbf{U}_\mathcal{H}\|_2 \leq \Big(&C_{\mathrm{LU}}h^{-1}\mathrm{depth}(\mathbb{T}_\mathcal{I})e^{-br^{1/(d+1)}} \\ &+ C_{\mathrm{LU}}^2 h^{-2}\mathrm{depth}(\mathbb{T}_\mathcal{I})^2 e^{-2br^{1/(d+1)}}\Big)\|\mathbf{A}\|_2, \end{aligned} \tag{38}$$

where $C_{\mathrm{LU}} = C_{\mathrm{sp}}C_{\mathrm{apx}}(\kappa_2(\mathbf{U}) + \kappa_2(\mathbf{L}))$*, with the constant* C_{apx} *from Theorem 1 and the spectral condition numbers* $\kappa_2(\mathbf{U})$*,* $\kappa_2(\mathbf{L})$*.*

Proof With Lemma 6, we get a low rank approximation of an admissible subblock $\tau'\times\sigma'$ of the lower triangular part of $\mathbf{L}(\tau)$ by

$$\begin{aligned}&\left\|\mathbf{S}(\tau,\sigma)\mathbf{U}(\sigma)^{-1}|_{\tau'\times\sigma'} - \mathbf{S}_\mathcal{H}(\tau',\sigma')\mathbf{U}(\sigma')^{-1}\right\|_2 \\ &= \left\|\mathbf{S}(\tau',\sigma')\mathbf{U}(\sigma')^{-1} - \mathbf{S}_\mathcal{H}(\tau',\sigma')\mathbf{U}(\sigma')^{-1}\right\|_2 \\ &\leq C_{\mathrm{apx}}h^{-1}e^{-br^{1/(d+1)}}\left\|\mathbf{U}(\sigma')^{-1}\right\|_2\|\mathbf{A}\|_2.\end{aligned}$$

Since $\mathbf{S}_\mathcal{H}(\tau',\sigma')\mathbf{U}(\sigma')^{-1}$ is a rank-r matrix, Lemma 5 immediately provides an $\mathcal{H}$-matrix approximation $\mathbf{L}_\mathcal{H}$ of the LU-factor $\mathbf{L}(\mathcal{I}) = \mathbf{L}$. Therefore, with Lemma 7 we get

$$\|\mathbf{L}-\mathbf{L}_\mathcal{H}\|_2 \leq C_{\mathrm{apx}}C_{\mathrm{sp}}h^{-1}\mathrm{depth}(\mathbb{T}_\mathcal{I})e^{-br^{1/(d+1)}}\left\|\mathbf{U}^{-1}\right\|_2\|\mathbf{A}\|_2$$

and in the same way an $\mathcal{H}$-matrix approximation $\mathbf{U}_\mathcal{H}$ of $\mathbf{U}(\mathcal{I}) = \mathbf{U}$ with

$$\|\mathbf{U}-\mathbf{U}_\mathcal{H}\|_2 \leq C_{\mathrm{apx}}C_{\mathrm{sp}}h^{-1}\mathrm{depth}(\mathbb{T}_\mathcal{I})e^{-br^{1/(d+1)}}\left\|\mathbf{L}^{-1}\right\|_2\|\mathbf{A}\|_2.$$

Since $\mathbf{A} = \mathbf{LU}$, the triangle inequality finally leads to

$$\begin{aligned}\|\mathbf{A}-\mathbf{L}_\mathcal{H}\mathbf{U}_\mathcal{H}\|_2 &\leq \|\mathbf{L}-\mathbf{L}_\mathcal{H}\|_2\|\mathbf{U}\|_2 + \|\mathbf{U}-\mathbf{U}_\mathcal{H}\|_2\|\mathbf{L}\|_2 + \|\mathbf{L}-\mathbf{L}_\mathcal{H}\|_2\|\mathbf{U}-\mathbf{U}_\mathcal{H}\|_2 \\ &\lesssim (\kappa_2(\mathbf{U})+\kappa_2(\mathbf{L}))\,\mathrm{depth}(\mathbb{T}_\mathcal{I})h^{-1}e^{-br^{1/(d+1)}}\|\mathbf{A}\|_2 \\ &\quad + \kappa_2(\mathbf{U})\kappa_2(\mathbf{L})\mathrm{depth}(\mathbb{T}_\mathcal{I})^2h^{-2}e^{-2br^{1/(d+1)}}\frac{\|\mathbf{A}\|_2^2}{\|\mathbf{L}\|_2\|\mathbf{U}\|_2},\end{aligned}$$

and the estimate $\|\mathbf{A}\|_2 \leq \|\mathbf{L}\|_2\|\mathbf{U}\|_2$ finishes the proof.

In the symmetric case, we may use the weaker admissibility condition (12) instead of (8) and obtain a result analogously to that of Theorem 4 for the Cholesky decomposition.

Corollary 2 *Let* $\mathbf{b} = \mathbf{0}$ *in* (1) *so that the resulting Galerkin matrix* $\mathbf{A}$ *is symmetric and positive definite. Let* $\mathbf{A} = \mathbf{C}\mathbf{C}^T$ *with* $\mathbf{C}$ *being a lower triangular matrix with positive diagonal entries* $\mathbf{C}_{jj} > 0$. *There exists a lower triangular blockwise rank-r matrix* $\mathbf{C}_{\mathcal{H}}$ *such that*

$$\left\| \mathbf{A} - \mathbf{C}_{\mathcal{H}}{\mathbf{C}_{\mathcal{H}}}^T \right\|_2 \leq \Big(C_{\mathrm{Ch}} h^{-1} \mathrm{depth}(\mathbb{T}_{\mathcal{I}}) e^{-br^{1/(d+1)}} + C_{\mathrm{Ch}}^2 h^{-2} \mathrm{depth}(\mathbb{T}_{\mathcal{I}})^2 e^{-2br^{1/(d+1)}} \Big) \left\| \mathbf{A} \right\|_2 , \tag{39}$$

where $C_{\mathrm{Ch}} = 2C_{\mathrm{sp}}C_{\mathrm{apx}}\sqrt{\kappa_2(\mathbf{A})}$, *with the constant* C_{apx} *from Theorem* 1 *and the spectral condition number* $\kappa_2(\mathbf{A})$.

Proof Since $\mathbf{A}$ is symmetric and positive definite, the Schur complements $\mathbf{S}(\tau, \tau)$ are symmetric and positive definite as well and therefore we get $\mathbf{D}(\tau)\mathbf{L}(\tau) = \mathbf{C}(\tau)$ in (37), where $\mathbf{D}(\tau)$ is a diagonal matrix, such that $\mathbf{D}(\tau)^2$ contains the diagonal elements of $\mathbf{U}(\tau)$. Moreover, we have $\|\mathbf{A}\|_2 = \|\mathbf{C}\|_2^2$ and $\kappa_2(\mathbf{C}) = \left\|\mathbf{C}^{-1}\right\|_2 \|\mathbf{C}\|_2 = \sqrt{\kappa_2(\mathbf{A})}$.

6 Numerical examples

In this section, we present numerical examples in two and three dimensions to confirm our theoretical estimates derived in the previous sections. Since numerical examples for the Dirichlet case have been studied before, e.g., in [4,20], we will focus on mixed Dirichlet–Neumann and pure Neumann problems in two and three dimensions.

With the choice $\eta = 2$ for the admissibility parameter in (8), the clustering is done by the standard geometric clustering algorithm, i.e., by splitting bounding boxes in half until they are admissible or smaller than the constant n_{leaf}, which we choose as $n_{\mathrm{leaf}} = 25$ for our computations. An approximation to the inverse Galerkin matrix is computed by using the bestapproximation via singular value decomposition. Throughout, we use the C-library HLiB [8] developed at the Max-Planck-Institute for Mathematics in the Sciences.

6.1 2D-diffusion

We consider the unit square $\Omega = (0, 1)^2$. The boundary $\Gamma = \partial\Omega$ is divided into the Neumann part $\Gamma_D := \{\mathbf{x} \in \Gamma \,:\, \mathbf{x}_1 = 0 \vee \mathbf{x}_2 = 0\}$ and the Dirichlet part $\Gamma_{\mathcal{N}} = \Gamma \backslash \overline{\Gamma_D}$. We consider the bilinear form $a(\cdot, \cdot) : H_0^1(\Omega, \Gamma_D) \times H_0^1(\Omega, \Gamma_D) \to \mathbb{R}$ corresponding to the mixed Dirichlet–Neumann Poisson problem

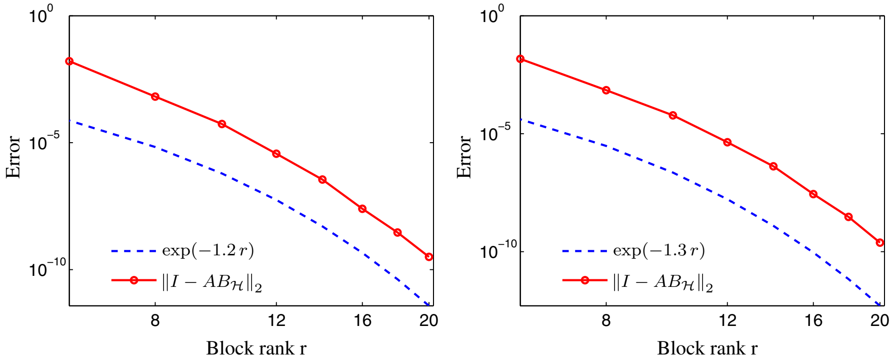

Fig. 1 Mixed boundary value problem (*left*), pure Neumann boundary value problem (*right*) in 2D

$$a(u, v) := \langle \nabla u, \nabla v \rangle_{L^2(\Omega)} \tag{40}$$

and use a lowest order Galerkin discretization in $S_0^{1,1}(\mathcal{T}_h, \Gamma_D)$.

As a second example, we study pure Neumann boundary conditions, i.e., $\Gamma = \Gamma_{\mathcal{N}}$, and use the bilinear form $a_{\mathcal{N}}(\cdot, \cdot) : H^1(\Omega) \times H^1(\Omega) \to \mathbb{R}$ corresponding to the stabilized Neumann Poisson problem

$$a_{\mathcal{N}}(u, v) := \langle \nabla u, \nabla v \rangle + \langle u, 1 \rangle \langle v, 1 \rangle \tag{41}$$

and a lowest order Galerkin discretization in $S^{1,1}(\mathcal{T}_h)$.

In Fig. 1, we compare the decrease of the upper bound $\|\mathbf{I} - \mathbf{A}\mathbf{B}_{\mathcal{H}}\|_2$ of the relative error with the increase in the block-rank for a fixed number $N = 262, 144$ of degrees of freedom, where the largest block of $\mathbf{B}_{\mathcal{H}}$ has a size of 32,768.

We observe exponential convergence in the block rank, where the convergence rate is $\exp(-br)$, which is even faster than the rate of $\exp(-br^{1/3})$ guaranteed by Theorem 2.

6.2 3D-diffusion

We consider the unit cube $\Omega = (0, 1)^3$ with the Dirichlet boundary $\Gamma_D := \{\mathbf{x} \in \Gamma : \exists i \in \{1, 2, 3\} : \mathbf{x}_i = 0\}$ and the Neumann part $\Gamma_{\mathcal{N}} = \Gamma \backslash \overline{\Gamma_D}$.

Again, we consider the bilinear forms (40) and (41) corresponding to the weak formulations of the Dirichlet–Neumann Poisson problem and the stabilized Neumann problem.

In Fig. 2, we compare the decrease of $\|\mathbf{I} - \mathbf{A}\mathbf{B}_{\mathcal{H}}\|_2$ with the increase in the block-rank for a fixed number $N = 32, 768$ of degrees of freedom, where the largest block of $\mathbf{B}_{\mathcal{H}}$ has a size of 4,096.

Comparing the results with our theoretical bound from Theorem 2, we empirically observe a rate of $e^{-br^{1/2}}$ instead of $e^{-br^{1/4}}$. Moreover, whether we study mixed boundary conditions or pure Neumann boundary conditions does not make any difference, as both model problems lead to similar computational results.

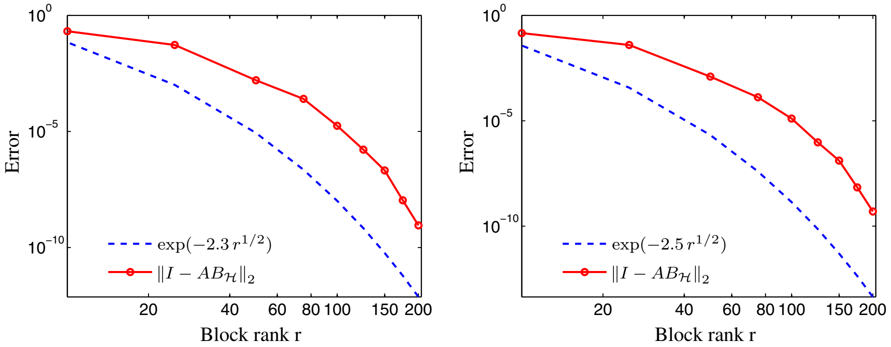

Fig. 2 Mixed boundary value problem (*left*), pure Neumann boundary value problem (*right*) in 3D

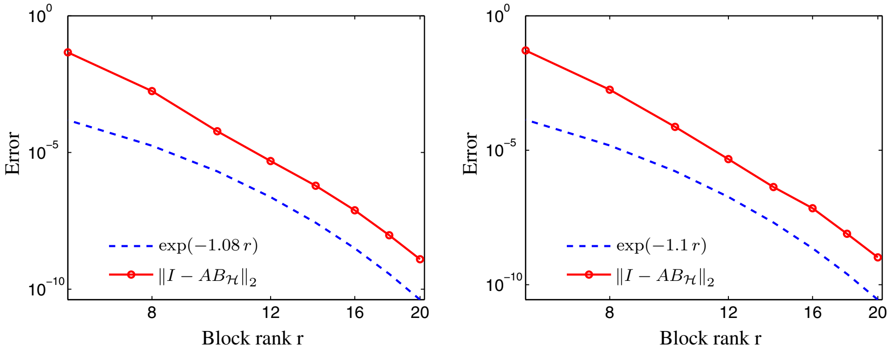

Fig. 3 2D convection-diffusion: mixed boundary value problem (*left*), pure Neumann boundary value problem (*right*)

6.3 Convection-diffusion

Finally, we study a convection-diffusion problem on the L-shaped domain $\Omega = (0,1) \times (0, \frac{1}{2}) \cup (0, \frac{1}{2}) \times [\frac{1}{2}, 1)$. The boundary $\Gamma = \partial\Omega$ is divided into the Neumann part $\Gamma_{\mathcal{N}} := \{\mathbf{x} \in \Gamma : \mathbf{x}_2 = 0 \vee \mathbf{x}_1 = 1\}$ and the Dirichlet part $\Gamma_D = \Gamma \backslash \overline{\Gamma_{\mathcal{N}}}$.

We consider the bilinear form $a(\cdot,\cdot) : H_0^1(\Omega, \Gamma_D) \times H_0^1(\Omega, \Gamma_D) \to \mathbb{R}$ corresponding to the mixed Dirichlet–Neumann Poisson problem

$$a(u, v) := c\, \langle \nabla u, \nabla v \rangle_{L^2(\Omega)} + \langle \mathbf{b} \cdot \nabla u, v \rangle_{L^2(\Omega)}$$

with $c = 10^{-2}$ and $\mathbf{b}(x_1, x_2) = (-x_2, x_1)^T$ and use a lowest order Galerkin discretization in $S_0^{1,1}(\mathcal{T}_h, \Gamma_D)$.

In Fig. 3, we observe exponential convergence of the upper bound $\|\mathbf{I} - \mathbf{A}\mathbf{B}_{\mathcal{H}}\|_2$ of the relative error with respect to the increase in the block-rank for a fixed number $N = 196,352$ of degrees of freedom, where the largest block of $\mathbf{B}_{\mathcal{H}}$ has a size of 24,544.

References

1. Adams, R.A.: Sobolev spaces, Pure and Applied Mathematics, vol. 65. Academic Press, New York (1975)
2. Bebendorf, M.: Hierarchical LU decomposition-based preconditioners for BEM. Computing **74**(3), 225–247 (2005)
3. Bebendorf, M.: Why finite element discretizations can be factored by triangular hierarchical matrices. SIAM J. Numer. Anal. **45**(4), 1472–1494 (2007)
4. Bebendorf, M., Hackbusch, W.: Existence of $\mathcal{H}$-matrix approximants to the inverse FE-matrix of elliptic operators with L^∞-coefficients. Numer. Math. **95**(1), 1–28 (2003)
5. Bennighof, J.K., Lehoucq, R.B.: An automated multilevel substructuring method for eigenspace computation in linear elastodynamics. SIAM J. Sci. Comput. **25**(6), 2084–2106 (electronic) (2004). doi:10.1137/S1064827502400650
6. Börm, S.: Approximation of solution operators of elliptic partial differential equations by $\mathcal{H}$- and $\mathcal{H}^2$-matrices. Numer. Math. **115**(2), 165–193 (2010)
7. Börm, S.: Efficient Numerical Methods for Non-local Operators, EMS Tracts in Mathematics, vol. 14. European Mathematical Society (EMS), Zürich (2010)
8. Börm, S., Grasedyck, L.: H-Lib—a library for $\mathcal{H}$-and $\mathcal{H}^2$-matrices. http://www.hlib.org (1999)
9. Brenner, S., Scott, L.: The Mathematical Theory of Finite Element Methods, Texts in Applied Mathematics, vol. 15. Springer, New York (2002)
10. Brenner, S.C.: The condition number of the Schur complement in domain decomposition. Numer. Math. **83**(2), 187–203 (1999)
11. Chandrasekaran, S., Dewilde, P., Gu, M., Somasunderam, N.: On the numerical rank of the off-diagonal blocks of Schur complements of discretized elliptic PDEs. SIAM J. Matrix Anal. Appl. **31**(5), 2261–2290 (2010). doi:10.1137/090775932
12. Dahmen, W., Faermann, B., Graham, I.G., Hackbusch, W., Sauter, S.A.: Inverse inequalities on non-quasiuniform meshes and application to the mortar element method. Math. Comput. **73**, 1107–1138 (2001)
13. Demkowicz, L., Kurtz, J., Pardo, D., Paszyński, M., Rachowicz, W., Zdunek, A.: Computing with *hp*-adaptive finite elements, vol. 2. Chapman & Hall/CRC Applied Mathematics and Nonlinear Science Series, Frontiers: three dimensional elliptic and Maxwell problems with applications. Chapman & Hall/CRC, Boca Raton (2008)
14. Ern, A., Guermond, J.L.: Evaluation of the condition number in linear systems arising in finite element approximations. M2AN Math. Model. Numer. Anal. **40**(1), 29–48 (2006)
15. Faustmann, M.: $\mathcal{H}$-matrix approximantion of inverses of FEM and BEM matrices. doctoral thesis, work in progress, Vienna (2015)
16. Faustmann, M., Melenk, J.M., Praetorius, D.: Existence of $\mathcal{H}$-matrix approximation to the inverse of BEM matrices: the simple layer operator. Institute for Analysis and Scientific Computing, Vienna University of Technology, Wien, Tech. Rep. in preparation (2013)
17. Faustmann, M., Melenk, J.M., Praetorius, D.: Existence of $\mathcal{H}$-matrix approximants to the inverses of BEM matrices: the hyper singular integral operator. Work in progress (2014)
18. Giebermann, K.: Multilevel approximation of boundary integral operators. Computing **67**(3), 183–207 (2001). doi:10.1007/s006070170005
19. Gillman, A., Martinsson, P.: A direct solver with $O(N)$ complexity for variable coefficient elliptic PDEs discretized via a high-order composite spectral collocation method. Tech. rep. (2013). arXiv:1302.5995 [math.NA]
20. Grasedyck, L.: Theorie und Anwendungen Hierarchischer Matrizen. Doctoral thesis, Kiel (2001) (in German)
21. Grasedyck, L.: Adaptive recompression of $\mathcal{H}$-matrices for BEM. Computing **74**(3), 205–223 (2005)
22. Grasedyck, L., Hackbusch, W.: Construction and arithmetics of $\mathcal{H}$-matrices. Computing **70**(4), 295–334 (2003)
23. Grasedyck, L., Hackbusch, W., Kriemann, R.: Performance of $\mathcal{H}$-LU preconditioning for sparse matrices. Comput. Methods Appl. Math. **8**(4), 336–349 (2008)
24. Grasedyck, L., Kriemann, R., Le Borne, S.: Parallel black box $\mathcal{H}$-LU preconditioning for elliptic boundary value problems. Comput. Vis. Sci. **11**(4–6), 273–291 (2008). doi:10.1007/s00791-008-0098-9
25. Grasedyck, L., Kriemann, R., Le Borne, S.: Domain decomposition based $\mathcal{H}$-LU preconditioning. Numer. Math. **112**(4), 565–600 (2009)

26. Greengard, L., Gueyffier, D., Martinsson, P.G., Rokhlin, V.: Fast direct solvers for integral equations in complex three-dimensional domains. Acta Numer. **18**, 243–275 (2009). doi:10.1017/S0962492906410011
27. Hackbusch, W.: A sparse matrix arithmetic based on $\mathcal{H}$-matrices. Introduction to $\mathcal{H}$-matrices. Computing **62**(2), 89–108 (1999)
28. Hackbusch, W.: Hierarchische Matrizen: Algorithmen und Analysis. Springer, Dordrecht (2009)
29. Hackbusch, W., Börm, S.: $\mathcal{H}^2$-matrix approximation of integral operators by interpolation. Appl. Numer. Math. 43(1–2), 129–143 (2002). doi:10.1016/S0168-9274(02)00121-6 (19th Dundee Biennial Conference on Numerical Analysis)
30. Hackbusch, W., Khoromskij, B., Sauter, S.A.: On $\mathcal{H}^2$-matrices. In: Lectures on applied mathematics (Munich, 1999), pp. 9–29. Springer, Berlin (2000)
31. Ho, K., Ying, L.: Hierarchical interpolative factorization for elliptic operators: differential equations. Tech. rep. (2013). arXiv:1307.2895 [math.NA]
32. Ho, K.L., Greengard, L.: A fast direct solver for structured linear systems by recursive skeletonization. SIAM J. Sci. Comput. **34**(5), A2507–A2532 (2012). doi:10.1137/120866683
33. Horn, R.A., Johnson, C.R.: Matrix Analysis, 2nd edn. Cambridge University Press, Cambridge (2013)
34. Karniadakis, G., Sherwin, S.: Spectral/hp Element Methods for CFD. Oxford University Press, Oxford (1999)
35. Le Borne, S., Grasedyck, L.: $\mathcal{H}$-matrix preconditioners in convection-dominated problems. SIAM J. Matrix Anal. Appl. **27**(4), 1172–1183 (electronic) (2006). doi:10.1137/040615845
36. Li, S., Gu, M., Wu, C.J., Xia, J.: New efficient and robust HSS Cholesky factorization of SPD matrices. SIAM J. Matrix Anal. Appl. **33**(3), 886–904 (2012). doi:10.1137/110851110
37. Lintner, M.: The eigenvalue problem for the 2D Laplacian in $\mathcal{H}$- matrix arithmetic and application to the heat and wave equation. Computing **72**(3–4), 293–323 (2004)
38. Martinsson, P.G.: A fast direct solver for a class of elliptic partial differential equations. J. Sci. Comput. **38**(3), 316–330 (2009). doi:10.1007/s10915-008-9240-6
39. Schmitz, P.G., Ying, L.: A fast direct solver for elliptic problems on general meshes in 2D. J. Comput. Phys. **231**(4), 1314–1338 (2012). doi:10.1016/j.jcp.2011.10.013
40. Schwab, C.: p- and hp-finite element methods. Theory and applications in solid and fluid mechanics. Numerical Mathematics and Scientific Computation. The Clarendon Press, Oxford University Press, New York (1998)
41. Scott, L.R., Zhang, S.: Finite element interpolation of nonsmooth functions satisfying boundary conditions. Math. Comput. **54**(190), 483–493 (1990)
42. Xia, J.: Efficient structured multifrontal factorization for general large sparse matrices. SIAM J. Sci. Comput. **35**(2), A832–A860 (2013). doi:10.1137/120867032
43. Xia, J., Chandrasekaran, S., Gu, M., Li, X.S.: Superfast multifrontal method for large structured linear systems of equations. SIAM J. Matrix Anal. Appl. **31**(3), 1382–1411 (2009). doi:10.1137/09074543X