MATHEMATICS OF COMPUTATION
Volume 85, Number 302, November 2016, Pages 2839–2861
http://dx.doi.org/10.1090/mcom/3086
Article electronically published on March 28, 2016

ON THE SPECTRAL EQUIVALENCE OF HIERARCHICAL MATRIX PRECONDITIONERS FOR ELLIPTIC PROBLEMS

M. BEBENDORF, M. BOLLHÖFER, AND M. BRATSCH

Abstract. We will discuss the spectral equivalence of hierarchical matrix approximations for second order elliptic problems. Our theory will show that a modified variant of the hierarchical matrix Cholesky decomposition which preserves test vectors while truncating blocks to lower rank will lead to a spectrally equivalent approximation when using an adapted truncation threshold. Our theory also covers the usual hierarchical Cholesky decomposition which does not preserve test vectors but expects a significantly more restrictive threshold adaption to obtain a spectrally equivalent approximation. Numerical experiments indicate that the adaption of the truncation parameter seems to be necessary for the traditional hierarchical Cholesky preconditioner to obtain mesh-independent convergence while the variant which preserves test vectors works in practice quite well even with a fixed parameter.

1. Introduction

Elliptic partial differential equations have often been in the focus of efficient numerical solution methods. Among many methods, multigrid methods [12, 17] have become a successful approach to treat these kinds of equations, though they suffer from some disadvantages, e.g., problems with anisotropies, non-smooth coefficients or irregular geometry. Algebraic multigrid methods (AMG) [16, 17] are often able to bypass these problems. More recently, a novel class of hierarchical matrix ($\mathcal{H}$-matrix) approximations [13, 15] has gained attention. These are based on a completely different approach, namely the admissibility of a hierarchy of subdomain pairs, which allow for a low-rank approximation. Among several hierarchical matrix approximations [5, 14], the $\mathcal{H}$-LU decomposition has turned out to be the most promising approach for preconditioning elliptic problems using $\mathcal{H}$-matrices. It is shown in [3, 4] that the hierarchical matrix approximation leads to an approximation of almost optimal complexity. However, for an optimal preconditioner, the accuracy of the $\mathcal{H}$-matrix has to be adapted to the mesh size h. This can be seen in numerical experiments (cf. e.g. [6] as well as Section 5), where one observes a dependence of the number of iteration steps on h whenever blocks are truncated to a fixed rank or even when using some constant relative tolerance ε for the truncation. The aim of this paper is to analyze and improve the technique such that in addition to an almost optimal complexity an optimal preconditioning effect is also achieved. In particular, our theory will show that the usual $\mathcal{H}$-matrix Cholesky preconditioning approach will lead to a bounded number of conjugate gradient steps, if the relative

Received by the editor August 13, 2014 and, in revised form, April 28, 2015.
2010 *Mathematics Subject Classification.* Primary 65F08, 65F50, 65N30.
Key words and phrases. Approximate LU decomposition, preconditioning, hierarchical matrices.
This work was supported by DFG collaborative research center SFB 611.

tolerance ε is reduced proportional to h^2. This increases the blockwise rank k, although k is known (see [4]) to grow only logarithmically as $h \to 0$. Numerical experiments indicate that the h-dependence of ε for the usual $\mathcal{H}$-matrix Cholesky decomposition is also necessary; that is why we believe that our result is sharp. Besides decreasing $\varepsilon \sim h^2$ globally, we will also show that for larger admissible blocks, $\varepsilon \sim h$ is already sufficient to obtain a spectrally equivalent preconditioner.

We have recently presented a modified $\mathcal{H}$-matrix Cholesky decomposition (see [6]) which locally preserves constant test vectors while truncating blocks to lower rank. Our theory in the current paper will also cover this modified approach. The results in Section 2.1 will show that when preserving constant test vectors, the smaller admissible blocks can be safely truncated with a relative tolerance ε that is independent of h, while only for some larger blocks the relative tolerance is required to shrink linearly with h. Since in practice there are not that many blocks of larger size, we regard this restriction as relatively mild. The main advantage of this modified approach, however, is that numerical experiments state that it stabilizes the matrix approximation. For instance, any perturbation applied to problems with small coercivity such as boundary value problems with dominating Neumann boundary conditions is likely to ruin important properties of the original discretisation (e.g., positivity) unless some kind of stabilization is applied; see [6] for a numerical comparison with AMG on such kind of problems.

The computation of $\mathcal{H}$-matrix approximations involves many truncation operations due to the fact that the sum of two rank-k matrices usually exceeds rank k. Truncations appear not only when two $\mathcal{H}$-matrices are added. Also, the product of two $\mathcal{H}$-matrices of the same block format usually results in a significantly different structure. Hence, the analysis of the actual $\mathcal{H}$-matrix LU factorization algorithm is cumbersome. Additionally, there are numerous ways to do the truncation in practice. We will therefore analyze the new techniques of this article for an algorithm that was used in [4] to show the existence of approximate $\mathcal{H}$-matrix LU decompositions. While this does not fully analyze the actual $\mathcal{H}$-matrix algorithm, it gives some mathematical underpinning to the success of the $\mathcal{H}$-matrix calculus when it is employed to compute approximate preconditioners.

The paper is organized as follows. Section 2 will establish a general relation between $\mathcal{H}$-matrix approximation and spectral equivalence. This will cover the usual truncation and also the truncation which preserves given vectors. These results will then be applied to approximations obtained from $\mathcal{H}$-matrix LU factorization in Section 3. From these results, preconditioners will be derived and their complexity will be analyzed in Section 4 and numerically illustrated in Section 5.

2. Hierarchical matrices and spectral equivalence

In the following, let $\Omega \subset \mathbb{R}^3$ be a bounded polyhedral domain. We confine ourselves to problems of the type

$$\mathcal{L}u = f \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega \tag{1}$$

with $\mathcal{L}u := -\operatorname{div} C\nabla u$, where $c_{ij} \in L^\infty(\Omega)$, $1 \le i, j \le 3$. The ellipticity of $\mathcal{L}$ is expressed by the assumption that for almost all $\xi \in \Omega$,

$$0 < \lambda_{\mathcal{L}} \le \lambda(\xi) \le \Lambda_{\mathcal{L}},$$

for all eigenvalues $\lambda(\xi)$ of the symmetric matrix $C(\xi) \in \mathbb{R}^{3\times 3}$.

Throughout this paper we will consider quasi-uniform triangulations $\mathcal{T}_h$ of the computational domain $\Omega \subset \mathbb{R}^3$. Furthermore, $V_h \subset H_0^1(\Omega)$ will refer to a conforming finite element space associated with our quasi-uniform triangulation, where

$$h := \max_{i \in I} \operatorname{diam} X_i,$$

$I := \{1, \ldots, n\}$, and $X_i := \operatorname{supp} \varphi_i$ denotes the supports of the piecewise linear finite element basis $\{\varphi_i\}_{i \in I}$ of V_h. For $t \subset I$ we define the support of t as $X_t := \bigcup_{i \in t} \operatorname{int} X_i$. We will make use of the natural injection $\mathcal{J}_t : \mathbb{R}^t \to V_h$ defined as

$$\mathcal{J}_t x = \sum_{i \in t} x_i \varphi_i,$$

which satisfies

$$c_{\mathcal{J}} \|x\|_2 \leq \sqrt{\frac{|t|}{|X_t|}} \, \|\mathcal{J}_t x\|_{L^2(X_t)} \leq c'_{\mathcal{J}} \|x\|_2, \quad x \in \mathbb{R}^t. \tag{2}$$

The mass matrix $M = (m_{ij})_{i,j \in I}$ is defined via

$$m_{ij} = (\varphi_i, \varphi_j)_{L^2(\Omega)}, \quad i, j \in I,$$

and satisfies

$$\|M^{1/2} x\|_2 = \|\mathcal{J}_I x\|_{L^2(\Omega)} \quad \text{for all } x \in \mathbb{R}^I. \tag{3}$$

It is well known that the mass matrix has a bounded condition number, i.e., there is $c_M > 0$ independent of h such that

$$\kappa(M) := \|M\|_2 \|M^{-1}\|_2 \leq c_M. \tag{4}$$

Furthermore, let $A \in \mathbb{R}^{I \times I}$ be the finite element stiffness matrix, i.e.,

$$a_{ij} = a(\varphi_j, \varphi_i), \quad i, j \in I,$$

where the bilinear form $a(u, v) := \int_\Omega \nabla v^T C \nabla u \, \mathrm{d}\xi$ refers to the weak form of the elliptic operator $\mathcal{L}$ and is therefore coercive, i.e.,

$$a(u, u) \geq \gamma \|\nabla u\|^2_{L^2(\Omega)} \quad \text{for all } u \in H_0^1(\Omega) \tag{5}$$

with some constant $\gamma > 0$. In particular, this shows that A is symmetric positive definite.

Many existing fast methods for the numerical solution of (1) are based on multilevel structures. The efficiency of $\mathcal{H}$-matrices is due to two principles, hierarchical matrix partitioning and low-rank representation. For an appropriate partition $\mathcal{P}$ of the set of matrix indices $I \times I$, a cluster tree T_I is constructed by recursively subdividing I. The subdivision of a cluster t is done using bisection such that indices which are in some sense close to each other are grouped together into the same cluster t_1 and t_2. The subdivision is continued as long as a cluster contains a minimum number $n_{\min}$ of indices. This yields the binary tree T_I. We denote by $L(T_I) - 1$ the height of the cluster tree T_I. Nodes with distance $\ell = 0, \ldots, L(T_I) - 1$ from the root define the set $T_I^{(\ell)}$ and are referred to as nodes of level ℓ. The set of leaves of T_I will be denoted by $\mathcal{L}(T_I)$.

In the following, we will expect the cluster trees to be balanced. This means that in three spatial dimensions there is a positive constant c_D such that for all $t \in T_I^{(\ell)}$, $\ell = 0, \ldots, L(T_I) - 1$,

$$2^{-\ell/3} / c_D \leq \operatorname{diam} X_t \leq c_D 2^{-\ell/3}, \tag{6}$$

i.e., after three successive levels, the diameter is essentially half as long. In particular, we have for the leaf clusters $\operatorname{diam} X_t \sim h$ and therefore

$$2^{-(L(T_I)-1)/3}/c_D \leq h \leq c_D 2^{-(L(T_I)-1)/3}. \tag{7}$$

There are three strategies which are commonly used to create the subdivision of the indices. Two of them (bounding boxes [10] and principal component analysis [5]) use grid information, whereas the method presented in [7] is based on the matrix graph of a sparse matrix.

The block cluster tree $T_{I\times I}$ is built by recursively subdividing $I \times I$. Each block $t \times s$ is subdivided into the sons $t' \times s'$, where t' and s' are taken from the lists of sons of t and s in T_I, respectively. In the following, $T_{I\times I}^{(\ell)}$ denotes the ℓ-th level of the block cluster tree $T_{I\times I}$. The recursion is done for a block $b := t \times s$ until it is small enough or satisfies the so-called *admissibility condition*:

$$\min\{\operatorname{diam} X_t, \operatorname{diam} X_s\} \leq \eta \operatorname{dist}(X_t, X_s). \tag{8}$$

The previous condition (8) guarantees that the restriction A_b of $A \in \mathbb{R}^{I\times I}$ can be approximated by a matrix of low rank; see [5]. All other blocks are small enough and stored as a dense matrix.

Remark. Since the first two subdivision techniques are based on convex enveloping sets (boxes and balls, respectively) for which the admissibility condition is checked, we may assume that each X_t is convex.

The set of leaves of the block cluster tree $T_{I\times I}$ constitutes the partition $\mathcal{P}$. The constructed partition has the property that for a given cluster $t \in T_I$ a constantly bounded number of blocks $t \times s$ appear in $\mathcal{P}$. Hence, the *sparsity constant*

$$c_{\mathrm{sp}} := \max_{t\in T_I} |\{s \subset I : t \times s \in \mathcal{P}\}|$$

is bounded independently of the size of I; see [11]. Notice that this implies that the number of blocks increases linearly with n. Due to the representation by low-rank matrices, each matrix block causes a computational cost that scales linearly with its dimensions. The set of hierarchical matrices on the partition $\mathcal{P}$ and blockwise rank k is then defined as

$$\mathcal{H}(\mathcal{P}, k) := \{A \in \mathbb{R}^{I\times I} : \operatorname{rank} A_b \leq k \text{ for all } b \in \mathcal{P}\}.$$

Elements of $\mathcal{H}(\mathcal{P}, k)$ provide data-sparse representations of fully populated matrices because the elements of this set can be stored with logarithmic-linear complexity. In addition to storing and multiplying $\mathcal{H}$-matrices efficiently by a vector, also higher (approximate) operations such as addition, multiplication, inversion, and LU factorization can be performed with logarithmic-linear complexity.

2.1. **Spectral equivalence and filtering.** In this section, we provide the main ingredients for the spectral equivalence of hierarchical matrix preconditioners. We state the results of this section in a quite general fashion since we believe that this framework holds for a wider class of applications than the hierarchical Cholesky decomposition treated in this article.

The $\mathcal{H}$-matrix Cholesky decomposition[1] leads to an approximation $\tilde{A} \in \mathbb{R}^{I\times I}$ of the original system matrix $A \in \mathbb{R}^{I\times I}$. Estimates on the resulting error $E := A - \tilde{A}$

[1]When we refer to the $\mathcal{H}$-Cholesky decomposition, we mean the Cholesky decomposition with blockwise truncation to lower rank in the sense of the hierarchical matrix algebra [14].

will be discussed in Section 3.1. Due to the hierarchical structure, the error E can be represented as

$$E = \sum_{\ell=0}^{L(T_I)-1} E_\ell, \tag{9}$$

where each part E_ℓ is the restriction of the blocks from $\mathcal{P}$ to those blocks from the ℓ-th level $\mathcal{P} \cap T_{I\times I}^{(\ell)}$. Thus E_ℓ can be written as

$$E_\ell = \sum_{b\in\mathcal{P}\cap T_{I\times I}^{(\ell)}} E_b \quad \in \mathbb{R}^{I\times I} \tag{10}$$

and here E_b refers to the error in a single block $b = t \times s$ extended by zeros. If A is symmetric and positive definite, then (provided it can be factorized in $\mathcal{H}$-matrix arithmetic) $\tilde{A} = \tilde{L}\tilde{L}^T$ is symmetric and positive definite, too. Due to the tensorial structure of E_ℓ, it holds that

$$\|E_\ell\|_2 \le c_{\mathrm{sp}} \max_{b\in\mathcal{P}\cap T_{I\times I}^{(\ell)}} \|E_b\|_2; \tag{11}$$

see [11] or Lemma 2.16 in [5].

The following lemma establishes the relation between preservation and conditioning, i.e., we prove spectral equivalence of A and $\tilde{A}$ provided the approximation $\tilde{A}$ is exact on a subspace in the sense that

$$P_\ell^T E_\ell P_\ell = E_\ell, \quad \ell = 0, \dots, L(T_I) - 1, \tag{12}$$

with a corresponding projection $P_\ell \in \mathbb{R}^{I\times I}$. Strictly speaking, we are going to bound the condition number

$$\kappa(\tilde{A}^{-1}A) := \frac{\lambda_{\max}(\tilde{A}^{-1}A)}{\lambda_{\min}(\tilde{A}^{-1}A)} = \|A^{-1/2}\tilde{A}A^{-1/2}\|_2 \, \|(A^{-1/2}\tilde{A}A^{-1/2})^{-1}\|_2, \tag{13}$$

which can be used to determine the convergence ratio of the preconditioned conjugate gradient method; cf. [1]. As an example, P_ℓ could simply be the identity, which corresponds to the standard $\mathcal{H}$-matrix Cholesky factorization. Another example are projectors which locally preserve row/column sums. We leave the precise choice of P_ℓ open at this point and will specify it later.

Lemma 2.1. *Let $A, \tilde{A} \in \mathbb{R}^{I\times I}$ be symmetric and positive definite such that $E = A - \tilde{A}$ is split as in* (9) *and satisfies* (12). *If there is a constant $0 \le \delta < 1$ (independent of h) such that*

$$\sum_{\ell=0}^{L(T_I)-1} \|P_\ell A^{-1} P_\ell^T\|_2 \|E_\ell\|_2 \le \delta,$$

then

$$\kappa(\tilde{A}^{-1}A) \le \frac{1+\delta}{1-\delta}.$$

Proof. Due to the levelwise decomposition $E = \sum_{\ell=0}^{L(T_I)-1} E_\ell$ of E, we obtain

$$\|I - A^{-1/2}\tilde{A}A^{-1/2}\|_2 = \|A^{-1/2}EA^{-1/2}\|_2 \le \sum_{\ell=0}^{L(T_I)-1} \|A^{-1/2}E_\ell A^{-1/2}\|_2$$
$$= \sum_{\ell=0}^{L(T_I)-1} \|A^{-1/2}P_\ell^T E_\ell P_\ell A^{-1/2}\|_2 \le \sum_{\ell=0}^{L(T_I)-1} \|E_\ell\|_2 \|P_\ell A^{-1} P_\ell^T\|_2 \le \delta.$$

Using (13) and the Neumann series, it follows that

$$\begin{aligned}\kappa(\tilde{A}^{-1}A) &= \lambda_{\max}(A^{-1/2}\tilde{A}A^{-1/2})\lambda_{\max}((A^{-1/2}\tilde{A}A^{-1/2})^{-1}) \\ &= \|A^{-1/2}\tilde{A}A^{-1/2}\|_2 \|(A^{-1/2}\tilde{A}A^{-1/2})^{-1}\|_2 \\ &\le \Big(1 + \|I - A^{-1/2}\tilde{A}A^{-1/2}\|_2\Big) \sum_{k=0}^{\infty} \|I - A^{-1/2}\tilde{A}A^{-1/2}\|_2^k \\ &= \frac{1 + \|I - A^{-1/2}\tilde{A}A^{-1/2}\|_2}{1 - \|I - A^{-1/2}\tilde{A}A^{-1/2}\|_2},\end{aligned}$$

which leads to the assertion. □

In the rest of this section, we are going to investigate the choices mentioned for P_ℓ in (12). The first choice is the identity, which is trivial but corresponds to the usual $\mathcal{H}$-matrix approximation.

Lemma 2.2. *Let $P_\ell = \mathrm{Id}$ and let the error $E = A - \tilde{A}$ be split as in* (9)*, where each E_ℓ is decomposed as in* (10)*. Then there exists a constant $c_I > 0$ such that for any $\delta < c_I$ and any error satisfying*

$$\sum_{\ell=0}^{L(T_I)-1} \max_{b \in \mathcal{P} \cap T_{I\times I}^{(\ell)}} \|E_b\|_2 \le \delta \|M\|_2$$

it holds that

$$\kappa(\tilde{A}^{-1}A) \le \frac{c_I + \delta}{c_I - \delta}.$$

Proof. Using Friedrichs' inequality

$$\|\mathcal{J}_I x\|_{L^2(\Omega)} \le c_F \operatorname{diam}\Omega \, \|\nabla \mathcal{J}_I x\|_{L^2(\Omega)}, \quad x \in \mathbb{R}^I,$$

it follows from (3), (5), and (4) that

$$\|A^{-1}\|_2 = \sup_{x\neq 0} \frac{\|x\|_2^2}{x^T A x} \le \|M^{-1}\|_2 \sup_{x\neq 0} \frac{\|\mathcal{J}_I x\|_{L^2(\Omega)}^2}{a(\mathcal{J}_I x, \mathcal{J}_I x)} \le \gamma^{-1}\|M^{-1}\|_2 \sup_{x\neq 0} \frac{\|\mathcal{J}_I x\|_{L^2(\Omega)}^2}{\|\nabla \mathcal{J}_I x\|_{L^2(\Omega)}^2}$$
$$\le \gamma^{-1}(c_F \operatorname{diam}\Omega)^2 \|M^{-1}\|_2 \le \frac{c_M (c_F \operatorname{diam}\Omega)^2}{\gamma \|M\|_2}.$$

With (11) we obtain

$$\|A^{-1}\|_2 \sum_{\ell=0}^{L(T_I)-1} \|E_\ell\|_2 \le c_{\mathrm{sp}} \|A^{-1}\|_2 \sum_{\ell=0}^{L(T_I)-1} \max_{b \in \mathcal{P} \cap T_{I\times I}^{(\ell)}} \|E_b\|_2 \le \frac{c_{\mathrm{sp}} c_M (c_F \operatorname{diam}\Omega)^2}{\gamma} \delta.$$

Setting $c_I := \gamma / \big(c_{\mathrm{sp}} c_M (c_F \operatorname{diam}\Omega)^2\big)$, the assertion follows from Lemma 2.1 using $\delta/c_I < 1$ instead of δ. □

The second choice for P_ℓ investigated in this article has already been proposed and investigated from a practical point of view in [6]. Assume that the approximation $\tilde{A}$ on each block $b = t \times s \in T^{(\ell)}_{I\times I} \cap \mathcal{P}$ preserves the vector $\mathbf{1}_s \in \mathbb{R}^I$ defined as

$$\mathbf{1}_s := \begin{cases} 1, & i \in s, \\ 0, & \text{else}, \end{cases}$$

i.e., E_b satisfies $E_b\mathbf{1}_s = 0$ and $E_b^T\mathbf{1}_t = 0$. Due to the locality of E_b, we even have $E_b\mathbf{1}_r = 0 = E_b^T\mathbf{1}_r$ for all $r \in T_I^{(\ell)}$. For P_ℓ we set

$$P_\ell := \mathrm{Id} - Q_\ell \in \mathbb{R}^{I\times I}, \quad \text{where} \quad Q_\ell := \sum_{r\in T_I^{(\ell)}} \frac{\mathbf{1}_r\mathbf{1}_r^T}{|r|}.$$

Then P_ℓ satisfies $E_\ell P_\ell = E_\ell$ as

$$E_\ell Q_\ell = \sum_{b\in T^{(\ell)}_{I\times I}\cap P} E_b Q_\ell = 0.$$

The symmetry of E_ℓ and P_ℓ implies that we also have $P_\ell E_\ell = E_\ell$.

Our aim is to prove a result similar to Lemma 2.2 for this particular choice of P_ℓ. To this end, we state the following two auxiliary lemmas which will prepare Lemma 2.5. The latter is analogous to Lemma 2.2 but uses $P_\ell = \mathrm{Id} - Q_\ell$ instead.

In the following lemma, we will require the condition,

$$\|\mathcal{J}_I P_\ell x\|_{L^2(\Omega)} \le \varepsilon_\ell \|\nabla \mathcal{J}_I x\|_{L^2(\Omega)},$$

which acts as a strengthened Friedrichs' inequality. The previous condition will be discussed in Lemma 2.4 afterwards.

Lemma 2.3. *If P_ℓ satisfies $\|\mathcal{J}_I P_\ell x\|_{L^2(\Omega)} \le \varepsilon_\ell \|\nabla \mathcal{J}_I x\|_{L^2(\Omega)}$ for all $x \in \mathbb{R}^I$, then*

$$\|P_\ell A^{-1} P_\ell^T\|_2 \le \frac{c_M \varepsilon_\ell^2}{\gamma \|M\|_2}.$$

Proof. From $\|P_\ell A^{-1} P_\ell^T\|_2 = \|P_\ell A^{-1/2}\|_2^2$, (5), and (4) we obtain

$$\begin{aligned} \|P_\ell A^{-1} P_\ell^T\|_2 &= \sup_{x\neq 0} \frac{\|P_\ell A^{-1/2} x\|_2^2}{\|x\|_2^2} = \sup_{y\neq 0} \frac{\|P_\ell y\|_2^2}{y^T A y} = \sup_{y\neq 0} \frac{\|P_\ell y\|_2^2}{a(\mathcal{J}_I y, \mathcal{J}_I y)} \\ &\le \gamma^{-1} \sup_{y\neq 0} \frac{\|P_\ell y\|_2^2}{\|\nabla \mathcal{J}_I y\|_{L^2}^2} \le \frac{\|M^{-1}\|_2}{\gamma} \sup_{y\neq 0} \frac{\|\mathcal{J}_I P_\ell y\|_{L^2}^2}{\|\nabla \mathcal{J}_I y\|_{L^2}^2} \le \frac{c_M \varepsilon_\ell^2}{\gamma \|M\|_2}. \end{aligned}$$

□

The next lemma yields an upper bound on ε_ℓ and is inspired by a similar argument used in aggregation-based multigrid methods [18]. For easier readability we set

$$D_\ell := \max_{t\in T_I^{(\ell)}} \operatorname{diam} X_t$$

for each $\ell = 0, \dots, L(T_I) - 1$.

Lemma 2.4. *Let $P_\ell = \mathrm{Id} - Q_\ell$. Then there exists a constant $c_A > 0$ such that*

$$\|\mathcal{J}_I P_\ell x\|_{L^2(\Omega)} \le c_A D_\ell \, \|\nabla \mathcal{J}_I x\|_{L^2(\Omega)}.$$

Proof. For $t \in T_I^{(\ell)}$ let

$$\mathring{X}_t := \{\xi \in X_t : (\mathcal{J}_I \mathbf{1}_t)(\xi) = 1\}.$$

Notice that the assumption that X_t is convex together with a sufficiently large minimum cluster size $n_{\min}$ implies that $\mathring{X}_t$ has a non-empty interior.

Let $\mathring{\mathcal{J}}_t^* : L^2(\Omega) \to \mathbb{R}^t$ be defined by

$$x^T \mathring{\mathcal{J}}_t^* v = (\mathcal{J}_t x, v)_{L^2(\mathring{X}_t)} \quad \text{for all } x \in \mathbb{R}^t,\ v \in L^2(\Omega).$$

The matrix $\mathring{M}_t := \mathring{\mathcal{J}}_t^* \mathcal{J}_t \in \mathbb{R}^{t \times t}$ has the entries

$$\mathring{m}_{ij} := (\varphi_i, \varphi_j)_{L^2(\mathring{X}_t)}, \quad i, j \in t,$$

and hence is the mass matrix restricted to $\mathring{X}_t$. The mass matrix for X_t will be denoted by M_t. Since mass matrices have constant condition numbers (cf. (4)) and norms of the order h^3, we immediately obtain

$$\begin{aligned}\|\mathcal{J}_t P_\ell x\|_{L^2(X_t)}^2 &= \|M_t^{1/2} P_\ell x\|_2^2 \le \|M_t\|_2 \|P_\ell x\|_2^2 \le \|M_t\|_2 \|\mathring{M}_t^{-1}\|_2 \|\mathring{M}_t^{1/2} P_\ell x\|_2^2 \\ &= \|M_t\|_2 \|\mathring{M}_t^{-1}\|_2 \|\mathcal{J}_t P_\ell x\|_{L^2(\mathring{X}_t)}^2 \le c' \|\mathcal{J}_t P_\ell x\|_{L^2(\mathring{X}_t)}^2\end{aligned}$$

with some constant $c' > 0$ independent of h. Observe that

$$\mathcal{J}_t P_\ell x = \mathcal{J}_t x_t - \sum_{r \in T_I^{(\ell)}} \frac{\mathbf{1}_r^T x_r}{|r|} \mathcal{J}_t \mathbf{1}_r = \mathcal{J}_t x_t - \frac{\mathbf{1}_t^T x_t}{|t|} \mathcal{J}_t \mathbf{1}_t.$$

We define the linear functional $F : L^2(\Omega) \to \mathbb{R}$ as $F(v) := \frac{1}{|t|} \mathbf{1}_t^T \mathring{M}_t^{-1} \mathring{\mathcal{J}}_t^* v$. Then

$$\|F(\mathcal{J}_t x_t)\|_{L^2(\mathring{X}_t)} = \frac{\sqrt{|\mathring{X}_t|}}{|t|} |\mathbf{1}_t^T x_t| \le \sqrt{\frac{|\mathring{X}_t|}{|t|}} \|x_t\|_2 \le c_{\mathcal{J}} \|\mathcal{J}_t x_t\|_{L^2(X_t)},$$

where we have exploited (2). Since for all constant functions c it holds that

$$x^T \mathring{M}_t^{-1} \mathring{\mathcal{J}}_t^* c = (\mathcal{J}_t \mathring{M}_t^{-1} x, c)_{L^2(\mathring{X}_t)} = c (\mathcal{J}_t \mathring{M}_t^{-1} x, \mathcal{J}_t \mathbf{1}_t)_{L^2(\mathring{X}_t)} = c\, x^T \mathbf{1}_t,$$

we have $\mathring{M}_t^{-1} \mathring{\mathcal{J}}_t^* c = c\, \mathbf{1}_t$ and thus $F(c) = c$. Let $c_t := |X_t|^{-1} \int_{X_t} \mathcal{J}_t x_t \, \mathrm{d}\xi \in \mathbb{R}$ be the average of $\mathcal{J}_t x_t$ in X_t. Then Poincaré's inequality (see [2] for a Poincaré inequality with domain-independent constant in the case of convex domains) can be applied to X_t:

$$\begin{aligned}\|\mathcal{J}_t x_t - \frac{\mathbf{1}_t^T x_t}{|t|} \mathcal{J}_t \mathbf{1}_t\|_{L^2(\mathring{X}_t)} &= \|\mathcal{J}_t x_t - F(\mathcal{J}_t x_t)\|_{L^2(\mathring{X}_t)} \\ &= \|\mathcal{J}_t x_t - c_t - F(\mathcal{J}_t x_t - c_t)\|_{L^2(\mathring{X}_t)} \\ &\le (1 + c_{\mathcal{J}}) \|\mathcal{J}_t x_t - c_t\|_{L^2(X_t)} \\ &\le (1 + c_{\mathcal{J}})\, c_P \operatorname{diam} X_t \, \|\nabla \mathcal{J}_t x_t\|_{L^2(X_t)} \\ &= \hat{c}_P \operatorname{diam} X_t \, \|\nabla \mathcal{J}_I x\|_{L^2(X_t)}.\end{aligned}$$

This eventually leads to the desired bound

$$\begin{aligned}\|\mathcal{J}_I P_\ell x\|^2_{L^2(\Omega)} &\le \sum_{t\in T_I^{(\ell)}} \|\mathcal{J}_t P_\ell x\|^2_{L^2(X_t)} \le c' \sum_{t\in T_I^{(\ell)}} \|\mathcal{J}_t P_\ell x\|^2_{L^2(\mathring{X}_t)} \\ &= c' \sum_{t\in T_I^{(\ell)}} \|\mathcal{J}_t x_t - \frac{\mathbf{1}_t^T x_t}{|t|}\mathcal{J}_t \mathbf{1}_t\|^2_{L^2(\mathring{X}_t)} \\ &\le c'\hat{c}_P^2 \sum_{t\in T_I^{(\ell)}} (\operatorname{diam} X_t)^2 \|\nabla \mathcal{J}_I x\|^2_{L^2(X_t)} \\ &\le \underbrace{c''\hat{c}_P^2}_{=:c_A^2} D_\ell^2 \,\|\nabla \mathcal{J}_I x\|^2_{L^2(\Omega)}. \end{aligned}$$

□

Similar to Lemma 2.2, we are now going to analyze how the blockwise error influences the spectral equivalence when preserving some side constraint on each block.

Lemma 2.5. *Let $P_\ell = \mathrm{Id} - Q_\ell$ and let the error $E = A - \tilde{A}$ be split as in* (9)*, where each E_ℓ is decomposed as in* (10)*. Then there exists a constant $c_Q > 0$ such that for any $\delta < c_Q$ and any error E satisfying*

$$\sum_{\ell=0}^{L(T_I)-1} D_\ell^2 \max_{b\in\mathcal{P}\cap T_{I\times I}^{(\ell)}} \|E_b\|_2 \le \delta \|M\|_2$$

it holds that

$$\kappa(\tilde{A}^{-1}A) \le \frac{c_Q + \delta}{c_Q - \delta}.$$

Proof. Using Lemma 2.3 and Lemma 2.4, it follows that

$$\|P_\ell A^{-1} P_\ell\|_2 \le \frac{c_M c_A^2}{\gamma \|M\|_2} D_\ell^2.$$

Hence, applying (11), we obtain

$$\begin{aligned}\sum_{\ell=0}^{L(T_I)-1} \|P_\ell A^{-1} P_\ell\|_2 \|E_\ell\|_2 &\le c_{\mathrm{sp}} \sum_{\ell=0}^{L(T_I)-1} \|P_\ell A^{-1} P_\ell\|_2 \max_{b\in\mathcal{P}\cap T_{I\times I}^{(\ell)}} \|E_b\|_2 \\ &\le \frac{c_{\mathrm{sp}} c_M c_A^2}{\gamma \|M\|_2} \sum_{\ell=0}^{L(T_I)-1} D_\ell^2 \max_{b\in\mathcal{P}\cap T_{I\times I}^{(\ell)}} \|E_b\|_2 \le \frac{c_{\mathrm{sp}} c_M c_A^2}{\gamma}\delta.\end{aligned}$$

We set $c_Q := \gamma/(c_{\mathrm{sp}} c_M c_A^2)$. Then the assertion follows from Lemma 2.1 using δ/c_Q instead of δ. □

The two Lemmas 2.2 and 2.5 present criteria in which way the blockwise error has to be bounded to obtain a spectrally equivalent preconditioner with and without preservation of side constraints. Notice that the condition of Lemma 2.5 is weaker than the corresponding condition in Lemma 2.2 due to the additional factor $(\max_{t\in T_I^{(\ell)}} \operatorname{diam} X_t)^2$. The results of this section are applicable to any symmetric positive definite approximation $\tilde{A}$ of A subject to splitting the error $E = A - \tilde{A}$ according to (9) and (10). The next section shows how this can be used for the hierarchical Cholesky decomposition.

3. Spectrally equivalent preconditioners based on the $\mathcal{H}$-matrix Cholesky factorization

In this section we will show that the $\mathcal{H}$-matrix Cholesky decomposition $\tilde{A} = \tilde{L}\tilde{L}^T$ will lead to an error matrix $E = A - \tilde{A}$ in which any block E_b refers to the error that is obtained when block b is replaced by an approximate low-rank matrix during the factorization; cf. Section 3.1. This will then be used in Section 3.2 to translate the criteria from Lemma 2.2 and Lemma 2.5 to criteria for blockwise truncation errors in order to be able to modify the algorithms accordingly.

3.1. Errors caused by $\mathcal{H}$-matrix Cholesky factorization.

Due to its efficiency and robustness, the approximate $\mathcal{H}$-matrix Cholesky decomposition $\tilde{A} := \tilde{L}\tilde{L}^T$ is usually favored over the $\mathcal{H}$-matrix inverse of the finite element stiffness matrix A. In practice, the approximation error $E = A - \tilde{A}$ emerges from blockwise truncation of singular values of intermediate matrices. The set of matrix blocks on which truncation is performed depends on the matrix structure and even on the particular implementation of the approximate matrix operations. Hence, the analysis of the actual algorithm is cumbersome. To avoid technical difficulties that are likely to distract the reader from the real problem, in this section it will be analyzed how the overall error E is related with truncations performed during an $\mathcal{H}$-matrix Cholesky factorization algorithm that was used in [4] to show the existence of approximate $\mathcal{H}$-matrix LU decompositions.

The recursive construction of the usual Cholesky decompositions relies on the factorization

$$L_{tt}L_{tt}^T = S(t,t) \tag{14}$$

of Schur complements

$$S(t,s) := A_{ts} - A_{t\rho}A_{\rho\rho}^{-1}A_{\rho s},$$

where the index set ρ is defined as (see Figure 1)

$$\rho := \{i \in I : i < \min t \cup s\}.$$

In particular, $S(I,I) = A$ gives the Cholesky decomposition $A = LL^T$. The recursive construction of the subblock

$$L_{tt} = \begin{bmatrix} L_{t_1t_1} & \\ L_{t_2t_1} & L_{t_2t_2} \end{bmatrix}$$

of L follows from the identity (see [4])

$$S(t,t) = \begin{bmatrix} S(t_1,t_1) & S(t_1,t_2) \\ S(t_1,t_2)^T & S(t_2,t_2) - L_{t_2t_1}L_{t_2t_1}^T \end{bmatrix},$$

where t_1 and t_2 denote the sons of t, by the three factorizations

$$\begin{aligned} L_{t_1t_1}L_{t_1t_1}^T &= S(t_1,t_1), \\ L_{t_1t_1}L_{t_2t_1}^T &= S(t_1,t_2), \\ L_{t_2t_2}L_{t_2t_2}^T &= S(t_2,t_2) - L_{t_2t_1}L_{t_2t_1}^T. \end{aligned}$$

The first and the last factorization are of type (14), whereas the second is a forward substitution, which can also be expressed by a recursion.

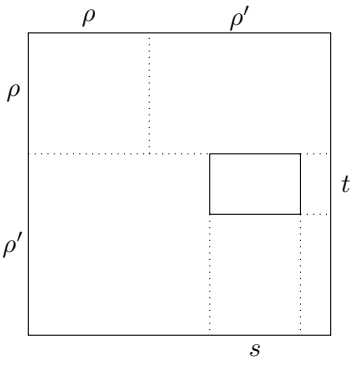

FIGURE 1. Index sets describing the Schur complement of a block $t \times s \in \mathcal{P}$.

Remark. We admit that this is not the standard way of defining the Cholesky decomposition, but it will turn out to be very useful in the following construction of low-rank approximations; see also [4]. For instance, assuming for a moment that $S(t_1, t_2)$ is rank r, the rank of $L_{t_2t_1}$ is also r. Notice that this is fundamentally different in incomplete LU factorizations, where the sparsity of $S(t_1, t_2)$ is not inherited by $L_{t_2t_1}$.

For an approximate recursive Cholesky factorization we have to address the problem of solving

$$\tilde{L}_{tt}\tilde{L}_{tt}^T + E_{tt} = \tilde{S}(t,t), \quad t \in T_I, \tag{15}$$

for a lower triangular matrix $\tilde{L}_{tt}$ and an error matrix E_{tt} instead of (14). The matrix $\tilde{S}(t,s) \in \mathbb{R}^{t\times s}$ is an approximation to the usual Schur complement $S(t,s)$ and will be constructed during the following recursive definition of the approximate Cholesky decomposition (15), which starts from the root I of T_I by setting

$$\tilde{S}(I,I) := A.$$

Due to the previous choice of $\tilde{S}(I,I)$, (15) will then yield the desired approximate Cholesky factorization

$$\tilde{L}\tilde{L}^T + E = A.$$

We just defined $\tilde{S}(I,I)$. Assume that $\tilde{S}(t,t)$ has already been defined for some $t \in T_I \setminus \mathcal{L}(T_I)$. Let $\tilde{S}(t_1,t_j) := \tilde{S}(t,t)|_{t_1t_j}$, $j = 1,2$, and define $\tilde{L}_{t_1t_1}$ recursively by the approximate Cholesky decomposition

$$\tilde{L}_{t_1t_1}\tilde{L}_{t_1t_1}^T + E_{t_1t_1} = \tilde{S}(t_1,t_1).$$

$\tilde{L}_{t_2t_1}$ is constructed via approximate forward substitution from

$$\tilde{L}_{t_1t_1}\tilde{L}_{t_2t_1}^T + E_{t_1t_2} = \tilde{S}(t_1,t_2). \tag{16}$$

With $\tilde{S}(t_2,t_2) := \tilde{S}(t,t)|_{t_2t_2} - \tilde{L}_{t_2t_1}\tilde{L}_{t_2t_1}^T$ the remaining subblock $\tilde{L}_{t_2t_2}$ of

$$\tilde{L}_{tt} = \begin{bmatrix} \tilde{L}_{t_1t_1} & \\ \tilde{L}_{t_2t_1} & \tilde{L}_{t_2t_2} \end{bmatrix}$$

is also constructed recursively by approximate Cholesky factorization

$$\tilde{L}_{t_2t_2}\tilde{L}_{t_2t_2}^T + E_{t_2t_2} = \tilde{S}(t_2,t_2).$$

Notice that this construction guarantees (15), where E_{tt} consists of the subblocks $E_{t_it_j}$, $i,j = 1,2$. If $t \in \mathcal{L}(T_I)$ is a leaf, then $\tilde{L}_{tt}$ is constructed via the pointwise Cholesky decomposition $\tilde{S}(t,t) = \tilde{L}_{tt}\tilde{L}_{tt}^T$ with $E_{tt} = 0$.

The forward substitution (16) has not yet been fully declared. For constructing $\tilde{L}_{st}$ from

$$\tilde{L}_{tt}\tilde{L}_{st}^T + E_{ts} = \tilde{S}(t,s), \quad t \times s \in T_{I\times I} \setminus \mathcal{P}, \tag{17}$$

let $\tilde{S}(t_1,s_j) := \tilde{S}(t,s)|_{t_1 s_j}$, $j = 1,2$, and recursively construct $\tilde{L}_{s_jt_1}$, $j = 1,2$, from the approximate forward substitution

$$\tilde{L}_{t_1t_1}\tilde{L}_{s_jt_1}^T + E_{t_1s_j} = \tilde{S}(t_1,s_j), \quad j = 1,2.$$

Furthermore, defining $\tilde{S}(t_2,s_j) := \tilde{S}(t,s)|_{t_2s_j} - \tilde{L}_{t_2t_1}\tilde{L}_{s_jt_1}^T$, we construct $\tilde{L}_{s_jt_2}$, $j = 1,2$, from

$$\tilde{L}_{t_2t_2}\tilde{L}_{s_jt_2}^T + E_{t_2s_j} = \tilde{S}(t_2,s_j), \quad j = 1,2.$$

Then

$$\tilde{L}_{st} = \begin{bmatrix} \tilde{L}_{s_1t_1} & \tilde{L}_{s_1t_2} \\ \tilde{L}_{s_2t_1} & \tilde{L}_{s_2t_2} \end{bmatrix}$$

satisfies (17). For leaf blocks $t \times s \in \mathcal{P}$ we approximate $\tilde{S}(t,s)$ with or without preservation of side constraints by a low-rank matrix $B_{ts} \in \mathbb{R}^{t\times s}$ such that

$$\|\tilde{S}(t,s) - B_{ts}\|_2 \le \varepsilon_\ell \|\tilde{S}(t,s)\|_2 \tag{18}$$

with a level-dependent accuracy $\varepsilon_\ell > 0$. The preservation of side constraints during this approximation can be done in a stable way via the Householder decomposition; see [6]. Then, we compute $\tilde{L}_{st}$ via the pointwise forward substitution, i.e., $\tilde{L}_{tt}\tilde{L}_{st}^T = B_{ts}$. As a consequence, (17) holds with $E_{ts} := \tilde{S}(t,s) - B_{ts}$.

Hence, approximation errors are introduced only during the forward substitution (17). In [4], it is shown that $\tilde{L} \in \mathcal{H}(\mathcal{P},k)$ is an $\mathcal{H}$-matrix with blockwise rank k depending polylogarithmically on the size of I, which proves the existence of $\mathcal{H}$-matrix approximations to the factors of the Cholesky decomposition.

Lemma 3.1. *The matrices resulting from the previous construction satisfy*

$$\tilde{S}(t,s) = A_{ts} - \tilde{L}_{t\rho}(\tilde{L}_{s\rho})^T.$$

Furthermore,

$$A_{ts} - (\tilde{L}\tilde{L}^T)_{ts} = E_{ts} = \tilde{S}(t,s) - B_{ts}$$

for all $t \times s \in \mathcal{P}$.

Proof. Let $t \times s \in \mathcal{P}$. The assertion follows from

$$\tilde{S}(t,s) = \tilde{L}_{tt}(\tilde{L}_{st})^T + E_{ts} = \tilde{L}_{tt}(\tilde{L}_{st})^T + A_{ts} - \tilde{A}_{ts} = A_{ts} - \tilde{L}_{t\rho}(\tilde{L}_{s\rho})^T$$

due to $\tilde{A}_{ts} = (\tilde{L}\tilde{L}^T)_{ts} = \tilde{L}_{t\rho}\tilde{L}_{s\rho}^T + \tilde{L}_{tt}\tilde{L}_{st}^T$. □

An important consequence of the previous lemma is that the error $E = A - \tilde{A}$ appearing in Lemma 2.2 and Lemma 2.5 for the choice $\tilde{A} = \tilde{L}\tilde{L}^T$ satisfies

$$\|E_{ts}\|_2 = \|\tilde{S}(t,s) - B_{ts}\|_2 \le \varepsilon_\ell \|\tilde{S}(t,s)\|_2, \quad t \times s \in \mathcal{P}; \tag{19}$$

see (18). Furthermore, the previous lemma shows that $\tilde{S}(t,s)$ can be viewed as a Schur complement containing the approximations made up to the block $t \times s$ in the

Cholesky factorization. Due to this interpretation, we assume that there exists a constant $c_S > 0$ such that

$$\|\tilde{S}(t,s)\|_2 \le c_S \|S(t,s)\|_2. \tag{20}$$

This stability assumption on the $\mathcal{H}$-matrix arithmetic is difficult to prove. It will not be investigated further in this paper but numerical experiments indicate that the norm of the approximate and exact Schur complement are close even for large ε_ℓ.

3.2. **Spectrally equivalent $\mathcal{H}$-matrix Cholesky preconditioners.** In the following, we are going to prove the main result (Theorems 3.5 and 3.6) of this article which compares preconditioners that are based on the $\mathcal{H}$-matrix Cholesky factorization with and without preservation of side constraints.

First of all, we discuss upper bounds on $\|S(t,s)\|_2$ for $t \times s \in \mathcal{P}$. Since $S(t,s)$ appears as part of some Schur complement of A, we immediately obtain

$$\|S(t,s)\|_2 \le \|A\|_2 \le ch^{-2}\|M\|_2.$$

To understand the interaction between the relative truncation strategy (18) that is used for the $\mathcal{H}$-matrix arithmetic and the error that is obtained for E_{ts}, we need a refined bound on the growth of $\|S(t,s)\|_2$ depending on the distance between X_t and X_s. The following theorem will be proved in Section 3.3.

Theorem 3.2. *The Schur complement of a block $t \times s \in \mathcal{P}$, $\operatorname{dist}(X_t, X_s) > 0$, is bounded by*

$$\|S(t,s)\|_2 \le \frac{c_a \|M\|_2}{h \operatorname{dist}(X_t, X_s)}$$

with some constant $c_a > 0$.

The previous theorem together with (19) and (20) leads to the following bound on the blockwise error introduced during $\mathcal{H}$-matrix arithmetic. Similar to the maximum diameter D_ℓ of clusters in level ℓ, we set

$$d_\ell := \min_{t \in T_I^{(\ell)}} \operatorname{diam} X_t.$$

Lemma 3.3. *Let assumption* (20) *be valid. Assume that the $\mathcal{H}$-matrix arithmetic is performed with a relative truncation as in* (18) *using a levelwise accuracy ε_ℓ. Then there exists a constant c_E such that*

$$\|E_b\|_2 \le c_E \frac{\varepsilon_\ell}{h\, d_\ell} \|M\|_2, \quad b \in \mathcal{P} \cap T_{I\times I}^{(\ell)}.$$

Proof. Using (19) and (20), it follows that

$$\|E_{ts}\|_2 \le c_S \varepsilon_\ell \|S(t,s)\|_2, \quad t \times s \in \mathcal{P} \cap T_{I\times I}^{(\ell)}.$$

Applying Theorem 3.2 and the admissibility condition (8), the blockwise error can be bounded in the following way,

$$\|E_{ts}\|_2 \le c_a c_S \frac{\varepsilon_\ell \|M\|_2}{h \operatorname{dist}(X_t, X_s)} \le c_a c_S \eta \frac{\varepsilon_\ell \|M\|_2}{h \min\{\operatorname{diam} X_t, \operatorname{diam} X_s\}},$$

and the assertion follows setting $c_E := c_a c_S \eta$. □

Lemma 3.3 shows that for small t, s (i.e. for large ℓ) such that $\operatorname{diam} X_t \sim h$, $\operatorname{diam} X_s \sim h$ we obtain errors on the order of $\|E_{ts}\|_2 \sim \varepsilon_\ell h^{-2}\|M\|_2 \sim \varepsilon_\ell \|A\|_2$, while for the largest admissible clusters t, s, where $\operatorname{diam} X_t$ and $\operatorname{diam} X_s$ are approximately constant, we have $\|E_{ts}\|_2 \sim \varepsilon_\ell h^{-1}\|M\|_2 \sim \varepsilon_\ell h\|A\|_2$.

In the next theorem, as a by-product of our theory, we estimate the global error resulting from the hierarchical Cholesky decomposition without preservation of side constraints. Note that this theorem improves a previous result

$$\|A - \tilde{L}\tilde{U}\|_2 \le c_1 \varepsilon L(T_I) h^{-2} \|L\|_2 \|U\|_2 \le c_2 \varepsilon L(T_I) h^{-4} \|M\|_2$$

presented in [4] by more than two orders of magnitude with respect to h.

Theorem 3.4. *Let $P_\ell = \mathrm{Id}$ and let (20) be valid. Assume that the $\mathcal{H}$-matrix Cholesky decomposition is computed using the relative truncation strategy (18) with a fixed accuracy $\varepsilon_\ell = \varepsilon$. Then there exists a constant $c_e > 0$ such that*

$$\|A - \tilde{L}\tilde{L}^T\|_2 \le c_e \varepsilon h^{-2} \|M\|_2.$$

Proof. It follows from Lemma 3.3 that

$$\|A - \tilde{L}\tilde{L}^T\|_2 \le c_{\mathrm{sp}} \sum_{\ell=0}^{L(T_I)-1} \max_{b \in \mathcal{P} \cap T_{I\times I}^{(\ell)}} \|E_b\|_2 \le c_{\mathrm{sp}} c_E \varepsilon \|M\|_2 h^{-1} \sum_{\ell=0}^{L(T_I)-1} d_\ell^{-1}.$$

Furthermore, one obtains with (6) and (7) that

$$\begin{aligned}\sum_{\ell=0}^{L(T_I)-1} d_\ell^{-1} &\le c_D \sum_{\ell=0}^{L(T_I)-1} 2^{\ell/3} = c_D 2^{(L(T_I)-1)/3} \sum_{\ell=0}^{L(T_I)-1} 2^{(\ell - L(T_I)+1)/3} \\ &= c_D 2^{(L(T_I)-1)/3} \sum_{\ell=0}^{L(T_I)-1} 2^{-\ell/3} \le 5 c_D 2^{(L(T_I)-1)/3} \le 5 c_D^2 h^{-1}.\end{aligned}$$

We therefore set $c_e := 5 c_{\mathrm{sp}} c_E c_D^2$. □

Remark. Theorem 3.4 remains valid for $P_\ell = \mathrm{Id} - Q_\ell$, because the preservation of side constraints does not change the quality of the approximation. Nevertheless, it will be seen in Theorem 3.6 that the condition number of the preconditioned system benefits from this advancement.

Now we will present the main theorem for the spectral equivalence of the usual $\mathcal{H}$-matrix Cholesky preconditioner. On one side it will give bounds on the condition number of the preconditioned system when a fixed relative accuracy $\varepsilon_\ell = \varepsilon$ is used for the relative truncation in (18). On the other side, it presents a strategy how to adaptively choose ε_ℓ to obtain spectral equivalence.

Theorem 3.5. *Let $P_\ell = \mathrm{Id}$ and let (20) be valid. Assume that the $\mathcal{H}$-matrix Cholesky decomposition is computed using the relative truncation strategy (18).*

$\mathcal{H}$`Chol`: Let $\varepsilon_\ell = \varepsilon$ be fixed. Then there exists a constant $c_\alpha > 0$ such that for any $\varepsilon < h^2/c_\alpha$ we have

$$\kappa\big((\tilde{L}\tilde{L}^T)^{-1} A\big) \le \frac{1 + c_\alpha \varepsilon h^{-2}}{1 - c_\alpha \varepsilon h^{-2}}.$$

$\mathcal{H}$`Chols`: *Let $\varepsilon_\ell = \varepsilon h d_\ell$. Then there exists a constant $c_\beta > 0$ such that for any $\varepsilon < 1/(c_\beta L(T_I))$ we have*

$$\kappa\big((\tilde{L}\tilde{L}^T)^{-1}A\big) \leq \frac{1 + c_\beta \varepsilon L(T_I)}{1 - c_\beta \varepsilon L(T_I)}.$$

Proof. It follows with Lemma 3.3 that

$$\sum_{\ell=0}^{L(T_I)-1} \max_{b \in \mathcal{P} \cap T_{I\times I}^{(\ell)}} \|E_b\|_2 \leq c_E h^{-1} \|M\|_2 \sum_{\ell=0}^{L(T_I)-1} \frac{\varepsilon_\ell}{d_\ell}. \tag{21}$$

For the constant choice $\varepsilon_\ell = \varepsilon$, the geometric series can be bounded as in the proof of Theorem 3.4 by $5c_D^2 h^{-1}$ such that we can apply Lemma 2.2 using $\delta := 5c_D^2 c_E h^{-2}\varepsilon$. The first assertion thus follows from Lemma 2.2 setting $c_\alpha := 5c_D^2 c_E c_I^{-1}$.

The second assertion follows from (21) and Lemma 2.2 since now we are able to directly use $\delta = c_E \varepsilon L(T_I)$, and we set $c_\beta := c_E c_I^{-1}$. □

Theorem 3.5 shows that for a constant approximation accuracy ε the condition number of the preconditioned system may behave like $\mathcal{O}(h^{-2})$. For a spectrally equivalent version we may practically ignore the contribution $L(T_I)$ in Theorem 3.5, since it only grows logarithmically with the system size. In this case the accuracy ε_ℓ of small blocks needs a rescaling of $\varepsilon_\ell \sim h^2$, whereas for the larger ones $\varepsilon_\ell \sim h$ is sufficient. Although Theorem 3.5 only provides upper bounds on the condition number of the preconditioned system, we believe that these bounds are relatively sharp as we will illustrate in Section 5.

Compared with Theorem 3.5 we are now able to establish improved bounds for the novel modified hierarchical Cholesky decomposition that additionally preserves side constraints.

Theorem 3.6. *Let $P_\ell = \mathrm{Id} - Q_\ell$ and let* (20) *be valid. Assume that the modified $\mathcal{H}$-matrix Cholesky decomposition that locally preserves side constraints is computed using the relative truncation strategy* (18).

M$\mathcal{H}$`Chol`: *Let $\varepsilon_\ell = \varepsilon$ be fixed. Then there exists a constant $c_\gamma > 0$ such that for any $\varepsilon < h/c_\gamma$ we have*

$$\kappa\big((\tilde{L}\tilde{L}^T)^{-1}A\big) \leq \frac{1 + c_\gamma \varepsilon h^{-1}}{1 - c_\gamma \varepsilon h^{-1}}.$$

M$\mathcal{H}$`Chols`: *Let $\varepsilon_\ell = \varepsilon h D_\ell^{-1}$. Then there exists a constant $c_\delta > 0$ such that for any $\varepsilon < 1/(c_\delta L(T_I))$ we have*

$$\kappa\big((\tilde{L}\tilde{L}^T)^{-1}A\big) \leq \frac{1 + c_\delta \varepsilon L(T_I)}{1 - c_\delta \varepsilon L(T_I)}.$$

Proof. Using Lemma 3.3 and (6), it follows that

$$\begin{aligned}\sum_{\ell=0}^{L(T_I)-1} D_\ell^2 \max_{b \in \mathcal{P} \cap T_{I\times I}^{(\ell)}} \|E_b\|_2 &\leq c_E h^{-1} \|M\|_2 \sum_{\ell=0}^{L(T_I)-1} \varepsilon_\ell \frac{D_\ell^2}{d_\ell} \\ &\leq c_E c_D^2 h^{-1} \|M\|_2 \sum_{\ell=0}^{L(T_I)-1} \varepsilon_\ell \, D_\ell.\end{aligned}$$

The second part of the assertion follows from Lemma 2.5 using $\delta := c_E c_D^2 L(T_I)\varepsilon$ and we may choose $c_\delta := c_E c_D^2 / c_Q$.

For the first choice $\varepsilon_\ell = \varepsilon$, a geometric series needs to be estimated. Using (6), one obtains the bound

$$\sum_{\ell=0}^{L(T_I)-1} D_\ell \le c_D \sum_{\ell=0}^{L(T_I)-1} 2^{-\ell/3} \le 5c_D.$$

Thus for the first assertion we may apply Lemma 2.5 with $\delta = 5c_E c_D^3 h^{-1}\varepsilon$. As a result of Lemma 2.5 we obtain the first part of of the assertion setting $c_\gamma := 5c_E c_D^3/c_Q$. □

For a constant approximation accuracy ε, Theorem 3.6 yields a bound for the preconditioner which is of one order of magnitude better (with respect to h) than the bound without the preservation of side constraints. Furthermore, for a spectrally equivalent version the accuracy of larger blocks has to be rescaled with $\varepsilon_\ell \sim h$. In contrast to this, the smaller blocks are already of sufficient accuracy using a constant ε. It is important to notice that the approximation accuracy needs to be adapted only to the level ℓ. This is why we regard $\varepsilon_\ell \sim h$ for upper levels as a technically uncomplicated condition. Like in the case of Theorem 3.5 we have proven upper bounds on the condition number in Theorem 3.6. However, these also seem to be numerically close to the behavior of the preconditioner in our numerical examples; cf. Section 5.

To simplify further discussions, we will define the following hierarchical matrix preconditioners. First of all, $\mathcal{H}$`Chol` and $\mathcal{H}$`Chols` denote the usual $\mathcal{H}$-matrix Cholesky decomposition. Their approximation accuracies are chosen constant or adaptively scaled with respect to level ℓ as proposed in Theorem 3.5. Second, M$\mathcal{H}$`Chol` and M$\mathcal{H}$`Chols` result from the modified $\mathcal{H}$-matrix Cholesky decomposition with the preservation of side constraints. The approximation accuracies are chosen constant or adaptively scaled per level as defined in Theorem 3.6.

3.3. **Proof of Theorem 3.2.** So far, estimates on the Schur complement used in the analysis of $\mathcal{H}$-matrices rely on bounds on the inverse of the stiffness matrix; see [5]. The approach used in this article is to directly bound the Schur complement as done, for instance, in the analysis of domain decomposition methods; see [8] and [9]. This leads to sharper estimates than the detour via the inverse.

Lemma 3.7. *Let $x \in \mathbb{R}^t$ and $y \in \mathbb{R}^s$. Then there is $\phi_h := \mathcal{J}_I(x - \tilde{x}) \in V_h$, $\tilde{x} \in \mathbb{R}^\rho$, such that*

$$x^T S(t,s) y = a(\phi_h, \mathcal{J}_I y)$$

and $a(\phi_h, \mathcal{J}_I z) = 0$ for all $z \in \mathbb{R}^\rho$.

Proof. Observe that

$$x^T S(t,s) y = x^T A_{ts} y - x^T A_{t\rho} A_{\rho\rho}^{-1} A_{\rho s} y = a(\mathcal{J}_I x, \mathcal{J}_I y) - a(\mathcal{J}_I \tilde{x}, \mathcal{J}_I y) = a(\phi_h, \mathcal{J}_I y),$$

where we set $\tilde{x} := A_{\rho\rho}^{-T} A_{t\rho}^T x \in \mathbb{R}^\rho$. Furthermore, for $z \in \mathbb{R}^\rho$,

$$a(\phi_h, \mathcal{J}_I z) = x^T A_{t\rho} z - \tilde{x}^T A_{\rho\rho} z = x^T A_{t\rho} z - x^T A_{t\rho} A_{\rho\rho}^{-1} A_{\rho\rho} z = 0. \qquad \square$$

Finally, we prove a discrete Caccioppoli inequality.

Lemma 3.8. *Let $\operatorname{dist}(X_t, X_s) > 0$. Then it holds that*

$$\|\nabla \phi_h\|_{L^2(X_\rho \cap X_s)} \le \frac{c_{\mathcal{L}}}{\operatorname{dist}(X_t, X_s)} \|\phi_h\|_{L^2(\Omega)}$$

with $c_{\mathcal{L}} > 0$ independent of h.

Proof. Let $\hat{s} := \{i \in I : 2\,\mathrm{dist}(X_i, X_s) \le \mathrm{dist}(X_t, X_s)\}$. Then $s \subset \hat{s}$ and $\mathrm{dist}(X_t, X_{\hat{s}}) > 0$. Define a discrete cut-off function $\eta_h \in V_h$ such that

$$\mathrm{supp}\,\eta_h \subset X_{\hat{s}}, \quad \eta_h|_{X_s} = 1, \quad \text{and} \quad \|\nabla\eta_h\|_\infty \le 2/\mathrm{dist}(X_s, \partial X_{\hat{s}}).$$

By $\mathfrak{I}_h : C(\Omega) \to V_h$ we denote the nodal interpolation operator. Due to $\Sigma := \mathrm{supp}\,\eta_h^2\phi_h \subset X_{\hat{s}} \cap X_\rho$, we have that $\mathfrak{I}_h(\eta_h^2\phi_h) \in \{\mathcal{J}_I x,\, x \in \mathbb{R}^\rho\}$. The discrete harmonicity of ϕ_h in X_ρ (see Lemma 3.7) implies

$$\begin{aligned}
\lambda_{\mathcal{L}}\|\nabla(\eta_h\phi_h)\|^2_{L^2(\Sigma)} &\le a(\eta_h\phi_h, \eta_h\phi_h) = a(\phi_h, \eta_h^2\phi_h) + \int_\Sigma \phi_h^2(\nabla\eta_h)^T C\nabla\eta_h\,\mathrm{d}\xi \\
&= a(\phi_h, \eta_h^2\phi_h - \mathfrak{I}_h(\eta_h^2\phi_h)) + \int_\Sigma \phi_h^2(\nabla\eta_h)^T C\nabla\eta_h\,\mathrm{d}\xi \\
&\le \Lambda_{\mathcal{L}}\|\nabla\phi_h\|_{L^2(\Sigma)}\|\nabla[\eta_h^2\phi_h - \mathfrak{I}_h(\eta_h^2\phi_h)]\|_{L^2(\Sigma)} + \frac{4\Lambda_{\mathcal{L}}}{\mathrm{dist}^2(X_s, \partial X_{\hat{s}})}\|\phi_h\|^2_{L^2(\Sigma)}.
\end{aligned}$$

It is known that

$$\|\nabla[\eta_h^2\phi_h - \mathfrak{I}_h(\eta_h^2\phi_h)]\|^2_{L^2(\Sigma)} \le (ch)^2 \sum_{\tau\subset\Sigma} \|D^2(\eta_h^2\phi_h)\|^2_{L^2(\tau)}.$$

From

$$\|D^2(\eta_h^2\phi_h)\|_{L^2(\tau)} = 2\||\nabla\eta_h|^2\phi_h + 2\eta_h\nabla\eta_h\cdot\nabla\phi_h\|_{L^2(\tau)} \le \frac{c}{\mathrm{dist}(X_s, \partial X_{\hat{s}})}\|\nabla(\eta_h\phi_h)\|_{L^2(\tau)}$$

it follows that

$$\begin{aligned}
\|\nabla[\eta_h^2\phi_h - \mathfrak{I}_h(\eta_h^2\phi_h)]\|^2_{L^2(\Sigma)} &\le \frac{c^2h^2}{\mathrm{dist}^2(X_s, \partial X_{\hat{s}})} \sum_{\tau\subset\Sigma} \|\nabla(\eta_h\phi_h)\|^2_{L^2(\tau)} \\
&\le \frac{c^2h^2}{\mathrm{dist}^2(X_s, \partial X_{\hat{s}})}\|\nabla(\eta_h\phi_h)\|^2_{L^2(\Sigma)}.
\end{aligned}$$

Hence, using the inverse inequality $\|\nabla\phi_h\|_{L^2(\Sigma)} \le c_I h^{-1}\|\phi_h\|_{L^2(\Sigma)}$ and the estimate $2ab \le \delta a^2 + b^2/\delta$ we obtain

$$\begin{aligned}
\lambda_{\mathcal{L}}\|\nabla(\eta_h\phi_h)\|^2_{L^2(\Sigma)} &\le \frac{c\Lambda_{\mathcal{L}}h}{\mathrm{dist}(X_s, \partial X_{\hat{s}})}\|\nabla\phi_h\|_{L^2(\Sigma)}\|\nabla(\eta_h\phi_h)\|_{L^2(\Sigma)} \\
&\quad + \frac{4\Lambda_{\mathcal{L}}}{\mathrm{dist}^2(X_s, \partial X_{\hat{s}})}\|\phi_h\|^2_{L^2(\Sigma)} \\
&\le \frac{cc_I\Lambda_{\mathcal{L}}}{\mathrm{dist}(X_s, \partial X_{\hat{s}})}\|\phi_h\|_{L^2(\Sigma)}\|\nabla(\eta_h\phi_h)\|_{L^2(\Sigma)} + \frac{4\Lambda_{\mathcal{L}}}{\mathrm{dist}^2(X_s, \partial X_{\hat{s}})}\|\phi_h\|^2_{L^2(\Sigma)} \\
&\le \left(\delta + \frac{4\Lambda_{\mathcal{L}}}{\mathrm{dist}^2(X_s, \partial X_{\hat{s}})}\right)\|\phi_h\|^2_{L^2(\Sigma)} + \frac{(cc_I\Lambda_{\mathcal{L}})^2}{4\delta\,\mathrm{dist}^2(X_s, \partial X_{\hat{s}})}\|\nabla(\eta_h\phi_h)\|^2_{L^2(\Sigma)}
\end{aligned}$$

for all $\delta > 0$. The choice $\delta := (2\lambda_{\mathcal{L}})^{-1}(cc_I\Lambda_{\mathcal{L}})^2/\mathrm{dist}^2(X_s, \partial X_{\hat{s}})$ leads to

$$\|\nabla\phi_h\|^2_{L^2(X_\rho\cap X_s)} \le \|\nabla(\eta_h\phi_h)\|^2_{L^2(\Sigma)} \le \frac{(cc_I\Lambda_{\mathcal{L}}/\lambda_{\mathcal{L}})^2 + 8\Lambda_{\mathcal{L}}/\lambda_{\mathcal{L}}}{\mathrm{dist}^2(X_s, \partial X_{\hat{s}})}\|\phi_h\|^2_{L^2(\Sigma)}.$$

The assertion follows from $\mathrm{dist}(X_s, \partial X_{\hat{s}}) \ge \frac{1}{2}\,\mathrm{dist}(X_t, X_s)$. □

We are now ready to prove Theorem 3.2.

Proof of Theorem 3.2. Let $y \in \mathbb{R}^s$. Since $\operatorname{supp}\phi_h \subset X_t \cup X_\rho$ and $\operatorname{supp}\mathcal{J}_I y \subset X_s$, using the inverse inequality and (2) we obtain

$$\begin{aligned} a(\phi_h, \mathcal{J}_I y) &\le \|C\nabla\phi_h\|_{L^2(X_\rho\cap X_s)}\|\nabla\mathcal{J}_I y\|_{L^2(X_\rho\cap X_s)} \\ &\le c_I \Lambda h^{-1}\|\nabla\phi_h\|_{L^2(X_\rho\cap X_s)}\|\mathcal{J}_I y\|_{L^2(X_\rho\cap X_s)} \\ &\le c_I c'_{\mathcal{J}} \Lambda h^{1/2}\|\nabla\phi_h\|_{L^2(X_\rho\cap X_s)}\|y\|_2. \end{aligned}$$

With the previous lemmas we find

$$\begin{aligned} \|S(t,s)\|_2 &= \sup_{x\in\mathbb{R}^t,\, y\in\mathbb{R}^s} \frac{x^T S(t,s) y}{\|x\|_2\|y\|_2} = \sup_{x\in\mathbb{R}^t,\, y\in\mathbb{R}^s} \frac{a(\phi_h, \mathcal{J}_I y)}{\|x\|_2\|y\|_2} \\ &\le c h^{1/2} \sup_{x\in\mathbb{R}^t} \frac{\|\nabla\phi_h\|_{L^2(X_\rho\cap X_s)}}{\|x\|_2} \le \frac{c'}{\operatorname{dist}(X_t, X_s)} h^{1/2} \sup_{x\in\mathbb{R}^t} \frac{\|\phi_h\|_{L^2(\Omega)}}{\|x\|_2}. \end{aligned}$$

The assertion follows from

$$\|\phi_h\|_{L^2(\Omega)} = \|\mathcal{J}_I(x-\tilde{x})\|_{L^2(\Omega)} \le c'_{\mathcal{J}} h^{3/2}\|x-\tilde{x}\|_2 \le c'_{\mathcal{J}} h^{3/2}(1 + \|A_{t\rho}A_{\rho\rho}^{-1}\|_2)\|x\|_2$$

due to (2) and $\|M\|_2 \sim h^3$. □

4. Complexity of $\mathcal{H}$-Cholesky preconditioners

Let $b \in \mathcal{P}$ and let the respective blockwise approximation accuracy ε_b be given. Then the blockwise rank k_b of an $\mathcal{H}$-Cholesky decomposition can be bounded by

$$k_b \in \mathcal{O}\big(L(T_I)^\alpha |\log \varepsilon_b|^\beta\big), \tag{22}$$

with constants $\alpha, \beta > 0$; see [5] for further details. Hence, the blockwise rank depends only logarithmically on the approximation accuracy. This allows us to adapt ε_b as proposed for the preconditioners $\mathcal{H}$`Chols` and `M`$\mathcal{H}$`Chols` without destroying the logarithmic-linear complexity of the $\mathcal{H}$-matrix arithmetic.

In the following theorem, we estimate the memory consumption and the computational complexity of the preconditioners $\mathcal{H}$`Chol` and `M`$\mathcal{H}$`Chol`.

Theorem 4.1. *The memory consumption of the preconditioners* $\mathcal{H}$`Chol` *and* `M`$\mathcal{H}$`Chol` *is of the order* $L(T_I)^{\alpha+1}|I|$*, the computational complexity is of the order* $L(T_I)^{2(\alpha+1)}|I|$.

Proof. The storage requirements of $\mathcal{H}$`Chol` are of the order $kL(T_I)|I|$, its computational complexity is of the order $k^2L^2(T_I)|I|$, where k denotes the maximum rank among the blocks; see [5]. Since the side constraints in `M`$\mathcal{H}$`Chol` increase the local rank by a constant, the overall complexity does not change. Hence, the assertion follows with (22). □

Theorem 4.1 shows that for sufficiently large matrices the preconditioner `M`$\mathcal{H}$`Chol` should be favored over $\mathcal{H}$`Chol` due to its better preconditioning properties; compare Theorem 3.5 and Theorem 3.6. In the following theorem, we estimate the memory consumption and the computational complexity of the spectrally equivalent preconditioners $\mathcal{H}$`Chols` and `M`$\mathcal{H}$`Chols`.

Theorem 4.2. *The memory consumption and the computational complexity of the preconditioners* $\mathcal{H}$`Chols` *and* `M`$\mathcal{H}$`Chols` *is of the order* $L(T_I)^{\alpha+\beta+1}|I|$ *and* $L(T_I)^{2(\alpha+\beta+1)}|I|$*, respectively.*

Proof. Using the balancedness of the cluster tree (7) and the dependence of the blockwise accuracy on h as in Theorem 3.5 and Theorem 3.6, we obtain for the preconditioners $\mathcal{H}$`Chols` and M$\mathcal{H}$`Chols` that

$$|\log \varepsilon_b| \lesssim |\log h| \sim |\log 2^{-L(T_I)/3}| \sim L(T_I), \quad b \in \mathcal{P}. \tag{23}$$

Hence, using (22) and (23) there exists a constant $c > 0$ such that

$$k_b \leq c\, L(T_I)^{\alpha+\beta}$$

and the assertion follows from the same arguments as in the proof of Theorem 4.1. □

Theorem 3.5 and Theorem 3.6 show that the preconditioner M$\mathcal{H}$`Chols` requires less adaption of the approximation accuracy on the respective block than $\mathcal{H}$`Chols` to obtain spectral equivalence. Still, we believe that the memory consumption of both is asymptotically the same and that Theorem 4.2 cannot be improved. Another important issue is the robustness of the preconditioner. Stability properties of $\mathcal{H}$`Chols` and M$\mathcal{H}$`Chols` will be investigated in the numerical results.

5. Numerical results

In this section, the results of Theorem 3.5 and Theorem 3.6 will be verified for an academic example. Afterwards, a more challenging example is considered to demonstrate the different stability properties of the spectrally equivalent preconditioners $\mathcal{H}$`Chols` and M$\mathcal{H}$`Chols`. Note that the focus of these tests is on the preconditioning properties rather than on the complexity of the preconditioner. The latter was analysed in Theorem 4.1 and Theorem 4.2 and its logarithmic-linear growth was observed in [6]. There, also a comparison with AMG can be found.

In the following tests, all linear systems were solved using the preconditioned conjugate gradient (PCG) method up to an accuracy of 10^{-10}. The numerical calculations were performed on a single core of an Intel Xeon X5482 processor at 3.2 GHz with 64 GB of core memory using the $\mathcal{H}$-matrix library A$\mathcal{H}$MED and the Intel Math Kernel Library (MKL) version 10.3.

Example 5.1. We consider the following boundary value problem

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega := (0,1)^3, \\ u &= 0 \quad \text{on } \partial\Omega. \end{aligned}$$

For the discretization linear ansatz functions have been chosen and the cluster trees of the different $\mathcal{H}$-matrices were constructed using a bounding box method. The minimal block size was set to $n_{\min} = 50$, and the admissibility parameter $\eta = 1.2$ was used. The truncation accuracy for the preconditioners $\mathcal{H}$`Chol` and M$\mathcal{H}$`Chol` was set to $\varepsilon = 0.1$. To obtain comparable results for the spectrally equivalent versions, $\mathcal{H}$`Chols` and M$\mathcal{H}$`Chols` were adapted so that the number of PCG steps is around 10 for the smallest test case with 3000 unknowns.

As can be seen from Table 1 and Table 2, $\mathcal{H}$`Chol` and M$\mathcal{H}$`Chol` approximate the stiffness matrix A in almost the same way but the preservation of side constraints leads to a significant reduction of the condition number. For $\mathcal{H}$`Chol` it can be seen that a factor eight in the number of unknowns n leads to a doubling in the number of PCG steps, which is in agreement with Theorem 3.5. Although the results in

TABLE 1. Preconditioners without preservation of side constraints for Example 5.1; notations $\star_1 := \|A-\tilde{A}\|_2/\|A\|_2$, $\star_2 :=$ PCG steps.

	$\mathcal{H}$Chol			$\mathcal{H}$Chols		
n	$\star_1$	$\kappa(\tilde{A}^{-1}A)$	$\star_2$	$\star_1$	$\kappa(\tilde{A}^{-1}A)$	$\star_2$
3 k	0,026	2,14	12	0,019	1,64	10
7 k	0,046	2,82	14	0,048	1,77	11
11 k	0,030	3,36	16	0,032	1,67	11
30 k	0,025	6,96	22	0,014	1,64	11
63 k	0,049	8,25	24	0,020	1,58	10
94 k	0,043	13,52	30	0,014	1,39	10
250 k	0,030	236,27	51	0,005	1,19	8

TABLE 2. Preconditioners with preservation of side constraints in Example 5.1; notations $\star_1 := \|A-\tilde{A}\|_2/\|A\|_2$, $\star_2 :=$ PCG steps.

	M$\mathcal{H}$Chol			M$\mathcal{H}$Chols		
n	$\star_1$	$\kappa(\tilde{A}^{-1}A)$	$\star_2$	$\star_1$	$\kappa(\tilde{A}^{-1}A)$	$\star_2$
3 k	0,025	1,26	9	0,021	1,26	9
7 k	0,043	1,31	9	0,043	1,30	9
11 k	0,027	1,31	10	0,025	1,28	9
30 k	0,025	1,32	10	0,025	1,31	10
63 k	0,048	1,38	10	0,047	1,36	10
95 k	0,044	1,65	12	0,040	1,58	11
250 k	0,027	1,41	11	0,027	1,35	10

Table 2 do not show significant changes in the condition number, we still believe that the estimate in Theorem 3.6 is sharp.

The almost constant number of PCG steps for the preconditioners $\mathcal{H}$Chols and M$\mathcal{H}$Chols is in accordance with the proposed levelwise adapted approximation accuracy in Theorem 3.5 and Theorem 3.6. An overview of the number of PCG steps used to solve the systems of linear equations with the different preconditioners is depicted in Figure 2.

Example 5.2. In the second example, we consider the diffusion problem

$$\begin{aligned} -\operatorname{div}(\sigma\nabla u) &= 0 \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \Gamma_1, \\ \frac{\partial u}{\partial \nu} &= I \quad \text{on } \Gamma_2, \\ \frac{\partial u}{\partial \nu} &= 0 \quad \text{on } \Gamma_3. \end{aligned}$$

The computational domain Ω is a conductor with the shape of a pyramid; see Figure 3. The boundary Γ_1 is the upper end of the conductor and Γ_2 is the lower end. We denote the remaining boundary as $\Gamma_3 := \partial\Omega\backslash(\Gamma_1\cup\Gamma_2)$. The conductivity σ of the conductor is set to $\sigma = 1$ in the left and $\sigma = 1000$ in the right half, while it is zero in the non-conductive part.

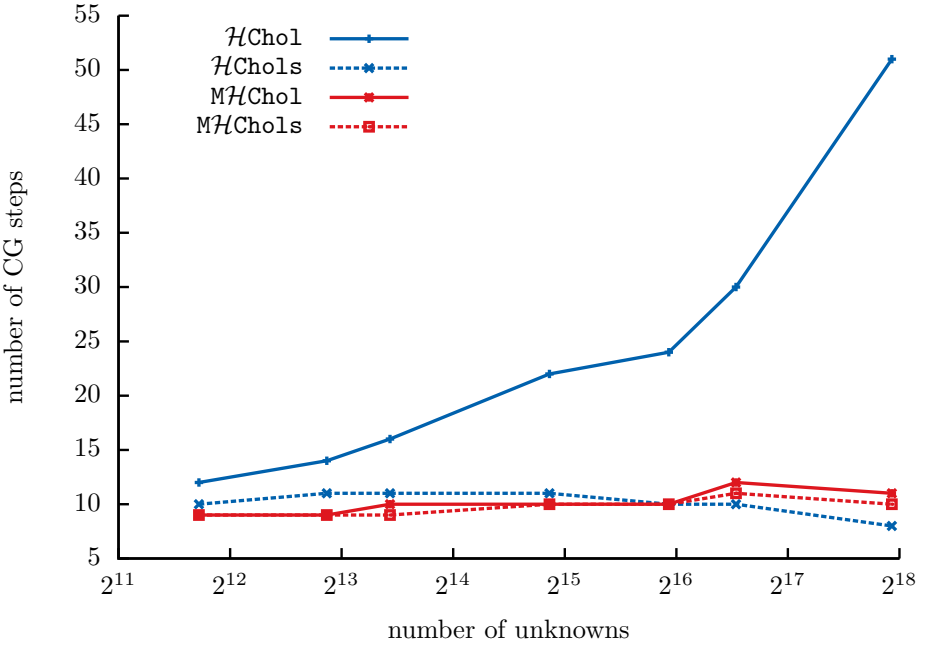

FIGURE 2. Comparison of the number of PCG steps for the considered preconditioners.

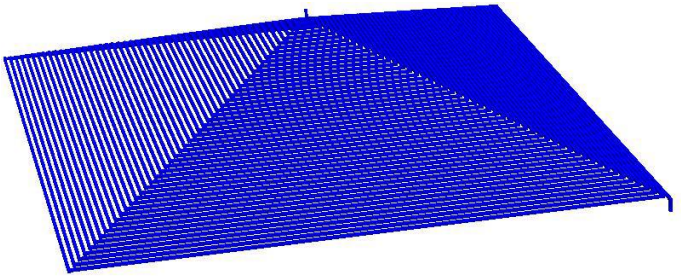

FIGURE 3. A conductor with the shape of a pyramid is the computational domain Ω of Example 2.

For the discretization, we used quadratic ansatz functions and the cluster trees were created using nested dissection; see [7]. The minimal blocksize was set to $n_{\min} = 150$, the admissibility parameter $\eta = 1.2$ was used.

The numerical results in Table 3 were obtained by employing $\mathcal{H}$`Chols` with the parameter $\varepsilon = 1e-3$ in Theorem 3.5. After computing $\mathcal{H}$`Chols` the accuracy for the preconditioner M$\mathcal{H}$`Chols` was adapted so that the resulting preconditioners have almost equal consumption of memory (difference is less than one percent). As can be seen from Table 3, it was not possible for the presented example to create a preconditioner $\mathcal{H}$`Chols` of similar size as M$\mathcal{H}$`Chols` for a large number of unknowns

TABLE 3. Spectrally equivalent preconditioners for Example 5.2.

n	memory	PCG steps $\mathcal{H}$Chols	PCG steps M$\mathcal{H}$Chols
66 k	72 MB	7	6
120 k	137 MB	14	10
390 k	596 MB	15	11
515 k	731 MB	n.a.	11
748 k	1 383 MB	n.a.	8

because the approximation without side constraints became indefinite. An explanation for this is that the preserved vectors span a linear space that approximate the small eigenvectors of the stiffness matrix. This leads to a stabilizing effect as discussed in [6] and makes the preconditioner M$\mathcal{H}$Chols practically more robust than $\mathcal{H}$Chols.

REFERENCES

[1] O. Axelsson, *Iterative Solution Methods*, Cambridge University Press, Cambridge, 1994. MR1276069 (95f:65005)

[2] M. Bebendorf, *A note on the Poincaré inequality for convex domains*, Z. Anal. Anwendungen **22** (2003), no. 4, 751–756, DOI 10.4171/ZAA/1170. MR2036927 (2004k:26025)

[3] M. Bebendorf, *Efficient inversion of the Galerkin matrix of general second-order elliptic operators with nonsmooth coefficients*, Math. Comp. **74** (2005), no. 251, 1179–1199 (electronic), DOI 10.1090/S0025-5718-04-01716-8. MR2136998 (2006b:65161)

[4] M. Bebendorf, *Why finite element discretizations can be factored by triangular hierarchical matrices*, SIAM J. Numer. Anal. **45** (2007), no. 4, 1472–1494 (electronic), DOI 10.1137/060669747. MR2338396 (2009f:65244)

[5] M. Bebendorf, *Hierarchical Matrices*, A Means to Efficiently Solve Elliptic Boundary Value Problems, Lecture Notes in Computational Science and Engineering, vol. 63, Springer-Verlag, Berlin, 2008. MR2451321 (2009k:15001)

[6] M. Bebendorf, M. Bollhöfer, and M. Bratsch, *Hierarchical matrix approximation with blockwise constraints*, BIT **53** (2013), no. 2, 311–339. MR3123848

[7] M. Bebendorf and T. Fischer, *On the purely algebraic data-sparse approximation of the inverse and the triangular factors of sparse matrices*, Numer. Linear Algebra Appl. **18** (2011), no. 1, 105–122, DOI 10.1002/nla.714. MR2769036 (2011k:65042)

[8] S. C. Brenner, *The condition number of the Schur complement in domain decomposition*, Numer. Math. **83** (1999), no. 2, 187–203, DOI 10.1007/s002110050446. MR1712684 (2000g:65114)

[9] M. Faustmann, J. M. Melenk, and D. Praetorius, *$\mathcal{H}$-matrix approximability of the inverse of FEM matrices*, Technical report, Institute for Analysis and Scientific Computing, 2013.

[10] K. Giebermann, *Multilevel approximation of boundary integral operators*, Computing **67** (2001), no. 3, 183–207, DOI 10.1007/s006070170005. MR1872653 (2002m:65128)

[11] L. Grasedyck and W. Hackbusch, *Construction and arithmetics of $\mathcal{H}$-matrices*, Computing **70** (2003), no. 4, 295–334, DOI 10.1007/s00607-003-0019-1. MR2011419 (2004i:65035)

[12] W. Hackbusch, *Multi-grid Methods and Applications*, volume 4 of *Springer Series in Computational Mathematics*. Springer, Berlin [u.a.], 1985.

[13] W. Hackbusch, *A sparse matrix arithmetic based on $\mathcal{H}$-matrices. I. Introduction to $\mathcal{H}$-matrices*, Computing **62** (1999), no. 2, 89–108, DOI 10.1007/s006070050015. MR1694265 (2000c:65039)

[14] W. Hackbusch. *Hierarchische Matrizen*. Springer-Verlag, 2009.

[15] W. Hackbusch and B. N. Khoromskij, *A sparse $\mathcal{H}$-matrix arithmetic. II. Application to multi-dimensional problems*, Computing **64** (2000), no. 1, 21–47. MR1755846 (2001i:65053)

[16] Y. Shapira, *Matrix-based Multigrid*, Theory and Applications, Numerical Methods and Algorithms, vol. 2, Kluwer Academic Publishers, Boston, MA, 2003. Theory and applications. MR2011771 (2004g:65002)

[17] U. Trottenberg, C.W. Oosterlee, and A. Schüller, *Multigrid: Basics, Parallelism and Adaptivity*, Academic Press, 2000.

[18] P. Vaněk, M. Brezina, and J. Mandel, *Convergence of algebraic multigrid based on smoothed aggregation*, Numer. Math. **88** (2001), no. 3, 559–579, DOI 10.1007/s211-001-8015-y. MR1835471 (2002c:65230)

(M. Bebendorf) Department of Mathematics, University of Bayreuth, Germany
E-mail address: mario.bebendorf@uni-bayreuth.de

(M. Bollhöfer) Institute for Computational Mathematics, TU Brunswick, Germany

(M. Bratsch) formerly Institute for Numerical Simulation, University of Bonn, Germany