Max-Planck-Institut
für Mathematik
in den Naturwissenschaften
Leipzig

Approximation of solution operators of elliptic partial differential equations by $\mathcal{H}$- and $\mathcal{H}^2$-matrices

(revised version: September 2007)

by

Steffen Börm

Preprint no.: 85 2007

Approximation of solution operators of elliptic partial differential equations by $\mathcal{H}$- and $\mathcal{H}^2$-matrices

Steffen Börm*

September 8, 2007

We investigate the problem of computing the inverses of stiffness matrices resulting from the finite element discretization of elliptic partial differential equations. Since the solution operators are non-local, the inverse matrices will in general be dense, therefore they cannot be represented by standard techniques. In this paper, we prove that these matrices can be approximated by $\mathcal{H}$- and $\mathcal{H}^2$-matrices. The key results are existence proofs for local low-rank approximations of the solution operator and its discrete counterpart, which give rise to error estimates for $\mathcal{H}$- and $\mathcal{H}^2$-matrix approximations of the entire matrices.

Keywords: Hierarchical matrices, data-sparse approximation, finite element methods, elliptic partial differential equations

AMS Subject Classification: 65F05, 65N30

Acknowledgement. A significant part of this work has been supported by the Institut für Geometrie und Praktische Mathematik of the RWTH Aachen.

1 Introduction

The discretization of a strongly elliptic partial differential equation by a standard finite element scheme leads to a linear system $Ax = b$ of equations that has to be solved in order to find the coefficients of the discrete approximation of the solution.

Systems of this type are usually solved by preconditioned Krylov or multilevel methods. For most non-multilevel preconditioners, the performance of Krylov methods deteriorates if the dimension of the linear system grows. Multilevel methods perform less than optimal when the coefficients of the differential operator are non-smooth or anisotropic.

*Max-Planck-Institut für Mathematik in den Naturwissenschaften, Inselstraße 22–26, 04103 Leipzig, Germany

The paper [2] introduces an alternative approach: the inverse of A is approximated by an $\mathcal{H}$-matrix, and it is possible to prove that the quality of the approximation depends only on the ratio of the maximal and minimal eigenvalues of the coefficient matrices, but not on their smoothness or the directions of anisotropy. Based on this result, the approximation of LU- and Cholesky-factorizations can be investigated [1, 12], and these approximative factorizations provide us with very efficient preconditioners for the original linear system.

The proof in [2] consists of four major steps: first a low-rank approximation of Green's function on a subdomain is derived using Cacciopoli's inequality and intermediate approximation steps by piecewise constant functions. In a second step, the integral operator corresponding to Green's function is discretized by Galerkin's method, giving rise to a dense matrix B. Due to the result of the first step, B can be approximated by an $\mathcal{H}$-matrix. In the third step, the connection between B and the inverse of A has to be established. Using the mass matrix M, it can be proven that $S := M^{-1}BM^{-1}$ corresponds to the L^2-projection of the solution operator into the discrete space. Since A^{-1} corresponds to the Galerkin projection into the discrete space, the Aubin-Nitsche lemma can be used to prove that S will converge to A^{-1} if the finite element grid is refined. In the last step of the proof, the inverse of the mass matrix M^{-1} is approximated by an $\mathcal{H}$-matrix, and a general result [11, Theorem 2.24] concerning the structure of the product of $\mathcal{H}$-matrices is used to prove that S can be approximated by a product of $\mathcal{H}$-matrix approximations of B and M^{-1}, although the rank of the product may be significantly larger than the rank of the original matrices.

In this paper, we improve the original result in two ways: first we approximate the solution operator directly instead of using an integral operator based on Green's function. This eliminates the second step of the original proof and yields error estimates with respect to the "natural" Sobolev norms instead of the weaker L^2-norm estimate given in the original paper.

More importantly, we replace the L^2-projections by Clément-type interpolation operators [8]. This means that we can construct an approximation of the inverse matrix directly without the detour via the inverse mass matrix M^{-1}, thus getting *local* blockwise estimates for the error instead of the global ones developed in [11].

The new approach also allows us to use the general framework of [4] to prove not only error estimates for $\mathcal{H}$-matrices, but also for the more efficient $\mathcal{H}^2$-matrices.

The paper is organized as follows: section 2 introduces the strongly elliptic and coercive model problem, section 3 proves that the corresponding solution operator can be approximated locally by low-rank operators, section 4 uses Clément-type interpolation operators to construct low-rank approximations of matrix blocks, and the sections 5 and 6 use these approximations to define $\mathcal{H}$- and $\mathcal{H}^2$-matrix approximation of A^{-1}.

2 Model problem

We fix a domain $\Omega \subseteq \mathbb{R}^d$ and a coefficient function $C : \Omega \to \mathbb{R}^{d\times d}$ satisfying

$$C(x) = C(x)^\top, \qquad \sigma(C(x)) \subseteq [\alpha, \beta] \qquad \text{for all } x \in \Omega.$$

We are interested in the partial differential operator

$$Lu := -\sum_{i,j=1}^{d} \partial_i C_{ij} \partial_j u \tag{1}$$

mapping $H_0^1(\Omega)$ into $H^{-1}(\Omega)$. For $f \in H^{-1}(\Omega)$, the partial differential equation

$$Lu = f$$

is equivalent to the variational equation

$$a(v,u) := \int_\Omega \langle \nabla v(x), C(x)\nabla u(x)\rangle_2 \, dx = f(v) \qquad \text{for all } v \in H_0^1(\Omega). \tag{2}$$

The bounds for the spectrum of C imply

$$\alpha\|w\|_2^2 \leq \langle C(x)w, w\rangle_2 = \|C(x)^{1/2}w\|_2^2 \leq \beta\|w\|_2^2, \qquad \text{for all } w \in \mathbb{R}^d.$$

Combining this inequality with the Cauchy-Schwarz inequality provides us with the upper bound

$$\begin{aligned}\langle \nabla v(x), C(x)\nabla u(x)\rangle_2 &= \langle C(x)^{1/2}\nabla v(x), C(x)^{1/2}\nabla u(x)\rangle_2 \\ &\leq \|C(x)^{1/2}\nabla v(x)\|_2 \|C(x)^{1/2}\nabla u(x)\|_2 \\ &\leq \beta\|\nabla v(x)\|_2\|\nabla u(x)\|_2,\end{aligned}$$

and the definition of the Sobolev space $H^1(\Omega)$ yields

$$|a(u,v)| \leq \beta\|u\|_{H^1(\Omega)}\|v\|_{H^1(\Omega)} \qquad \text{for all } u, v \in H^1(\Omega).$$

This means that the bilinear form a is bounded, i.e., continuous. Due to

$$\langle \nabla u(x), C(x)\nabla u(x)\rangle_2 \geq \alpha\|\nabla u(x)\|_2^2,$$

Friedrichs' inequality implies

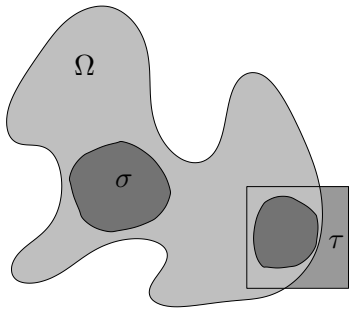
$$a(u,u) \geq \alpha\|\nabla u\|_{L^2(\Omega)}^2 \geq C_\Omega \alpha \|u\|_{H^1(\Omega)}^2 \qquad \text{for all } u \in H_0^1(\Omega),$$

i.e., a is a coercive bilinear form, therefore (2) and the equivalent (1) possess unique solutions [7].

3 Approximation of the solution operator

Let $\tau \subseteq \mathbb{R}^d$ be a convex set with $\tau \cap \Omega \neq \emptyset$. Let $\sigma \subseteq \Omega$ be a subset with $\operatorname{dist}(\tau, \sigma) > 0$.

Let $\epsilon \in \mathbb{R}_{>0}$. We are looking for a finite-dimensional space $V \subseteq H^1(\tau)$ such that for each right-hand side $f \in H^{-1}(\Omega)$ with $\operatorname{supp} f \subseteq \sigma$ the corresponding solution $u \in H_0^1(\Omega)$



of the variational equation (2) can be approximated in V, i.e., such that there exists a $v \in V$ with

$$\|u - v\|_{H^1(\tau)} \leq \epsilon \|f\|_{H^{-1}(\Omega)}.$$

Since V is required to be independent of f, this property implies that the interaction between the domains τ and σ can be described by a low-rank operator.

If the coefficient function C and the boundary of Ω were sufficiently smooth, interior regularity estimates would yield an estimate of the form

$$\|u\|_{H^m(\tau)} \leq C \operatorname{diam}(\tau)^m m! \|f\|_{H^{-1}(\Omega)}$$

and we could apply standard approximation theory to construct an approximating polynomial $\tilde{u}$.

In the general case, we have to use a refined approach first presented in [2]: since $u \in H^1(\Omega)$ holds, we can approximate the solution by a piecewise constant function, but the convergence rate will not be exponential. Projecting this function into a local space of L-harmonic functions (cf. Definition 1 below) yields an approximation v_1. We can apply a weak interior regularity argument to show that $v_1|_{\tau_1}$ is contained in $H^1(\tau_1)$ for a subset $\tau_1 \subseteq \Omega$, therefore the error $u_1 := u|_{\tau_1} - v_1|_{\tau_1}$ is also an L-harmonic function in $H^1(\tau_1)$, and the argument can be repeated until a sufficiently accurate approximation $v := v_1 + \ldots + v_p$ has been found.

The key element of the proof is the space of locally L-harmonic functions:

Definition 1 (Locally L-harmonic functions) *Let $\omega \subseteq \mathbb{R}^d$ be a domain (that may be unrelated to Ω). A function $u \in L^2(\omega)$ is called* locally L-harmonic *on ω if for all $K \subseteq \omega$ with $\operatorname{dist}(K, \partial\omega) > 0$ the following conditions hold:*

$$u|_K \in H^1(K), \tag{3a}$$

$$a(v, u|_\Omega) = 0 \qquad \text{for all } v \in H_0^1(\Omega) \text{ with } \operatorname{supp} v \subseteq K, \tag{3b}$$

$$u|_{\omega \setminus \Omega} = 0, \tag{3c}$$

The space of all locally L-harmonic functions on ω is denoted by $Z(\omega)$.

For functions in $Z(\omega)$, the following weak interior regularity estimate holds (cf. [2, Lemma 2.4]):

Lemma 2 (Cacciopoli inequality) *Let $u \in Z(\omega)$, and let $\tilde{\omega} \subseteq \omega$ be a domain with $\operatorname{dist}(\tilde{\omega}, \partial\omega) > 0$. Then we have $u|_{\tilde{\omega}} \in H^1(\tilde{\omega})$ and*

$$\|\nabla u\|_{L^2(\tilde{\omega})} \leq \frac{c_{\mathrm{reg}}}{\operatorname{dist}(\tilde{\omega}, \partial\omega)} \|u\|_{L^2(\omega)}, \qquad c_{\mathrm{reg}} := 4\sqrt{\beta/\alpha} \geq 4.$$

As mentioned before, we will use orthogonal projections to map functions from $L^2(\omega)$ into $Z(\omega)$. The construction of these projections is straightforward if $Z(\omega)$ is a complete set, i.e., closed in $L^2(\omega)$. Using Lemma 2, this property can be proven (cf. [2, Lemma 2.2]):

Lemma 3 *The space $Z(\omega)$ is a closed subspace of $L^2(\omega)$.*

We introduce the maximum-norm diameter

$$\begin{aligned} \operatorname{diam}_\infty(\omega) &:= \sup\{\|x - y\|_\infty \ : \ x, y \in \omega\} \\ &= \sup\{|x_i - y_i| \ : \ x, y \in \omega, i \in \{1, \ldots, d\}\} \end{aligned}$$

and can now state the basic approximation result (the proof is a slight modification of [2, Lemma 2.6]):

Lemma 4 (Finite-dimensional approximation) *Let $\omega \subseteq \mathbb{R}^d$ be a convex domain. Let $\ell \in \mathbb{N}$. Let Z be a closed subspace of $L^2(\omega)$. There is a space $V \subseteq Z$ with $\dim(V) \leq \ell^d$ such that for all $u \in Z \cap H^1(\omega)$ a function $v \in V$ can be found with*

$$\|u - v\|_{L^2(\omega)} \leq c_{\mathrm{apx}} \frac{\operatorname{diam}_\infty(\omega)}{\ell} \|\nabla u\|_{L^2(\omega)}, \qquad c_{\mathrm{apx}} := \frac{2\sqrt{d}}{\pi}.$$

Combining the construction of Lemma 4 with the regularity result of Lemma 4 allows us to find finite-dimensional spaces approximating the solutions of the variational equation (2):

Theorem 5 (Low-rank approximation) *Let $\eta \in \mathbb{R}_{>0}$ and $q \in (0, 1)$. There are constants $C_{\mathrm{apx}}, C_{\dim} \in \mathbb{R}_{>0}$ such that for all convex domains $\tau \subseteq \mathbb{R}^d$ and all $p \in \mathbb{N}_{\geq 2}$, we can find a space $V \subseteq L^2(\tau \cap \Omega)$ satisfying*

$$\dim V \leq C_{\dim} p^{d+1} \tag{4}$$

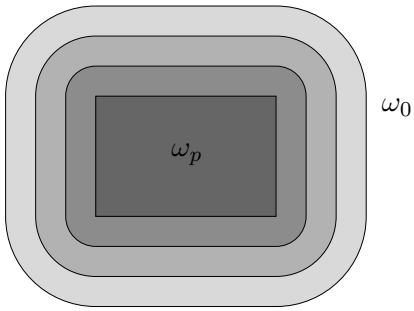
and for all domains $\sigma \subseteq \Omega$ with

$$\operatorname{diam}(\tau) \leq 2\eta \operatorname{dist}(\tau, \sigma) \tag{5}$$

and all right-hand sides $f \in H^{-1}(\Omega)$ with $\operatorname{supp} f \subseteq \sigma$, the corresponding solution $u \in H^1_0(\Omega)$ of the variational equation (2) can be approximated by a function $v \in V$ with

$$\|\nabla u|_\tau - \nabla v\|_{L^2(\tau)} \leq C_{\mathrm{apx}} q^p \|f\|_{H^{-1}(\Omega)}, \tag{6}$$

$$\|u|_\tau - v\|_{H^1(\tau)} \leq C_{\mathrm{apx}} (\operatorname{dist}(\tau, \sigma)/8 + 1) q^p \|f\|_{H^{-1}(\Omega)}. \tag{7}$$



Proof: Let $\tau \subseteq \mathbb{R}^d$ be a convex domain, let $\sigma \subseteq \Omega$ be a domain satisfying (5), let $\delta := \operatorname{dist}(\tau, \sigma)$, and let $p, \ell \in \mathbb{N}$. We introduce the domains

$$\omega_i := \left\{ x \in \mathbb{R}^d \ : \ \operatorname{dist}(x, \tau) \leq \frac{(p-i)\delta}{p} \right\} \qquad \text{for all } i \in \{0, \ldots, p\}.$$

In order to apply Lemma 2, we need an estimate for the distance between the boundaries of these subdomains. Let $i \in \{1, \ldots, p\}$, $x \in \omega_i$ and $y \in \partial\omega_{i-1}$. Due to $\operatorname{dist}(y, \tau) = (p-i+1)\delta/p$, we can find $z \in \tau$ with $\|y - z\|_2 \geq (p-i+1)\delta/p - \epsilon$ for all $\epsilon \in \mathbb{R}_{>0}$ and conclude

$$\|x - y\|_2 \geq \|y - z\|_2 - \|x - z\|_2 \geq \frac{(p-i+1)\delta}{p} - \epsilon - \frac{(p-i)\delta}{p} = \frac{\delta}{p} - \epsilon,$$

i.e., $\operatorname{dist}(\omega_i, \partial\omega_{i-1}) \geq \delta/p$.

Let $f \in H^{-1}(\Omega)$ with $\operatorname{supp} f \subseteq \sigma$. Let $u \in H_0^1(\Omega)$ be the corresponding solution of the variational equation (2). Extending u by zero if necessary yields $u_0 \in H_0^1(\Omega \cap \omega_0)$.

For all domains $K \subseteq \omega_0$ with $\operatorname{dist}(K, \partial\omega_0) > 0$, we have $u_0|_K \in H^1(K)$, and for all $v \in H_0^1(\Omega)$ with $\operatorname{supp} v \subseteq K$, we have $\operatorname{supp} v \cap \operatorname{supp} f = \emptyset$ and therefore

$$a(v, u_0|_\Omega) = a(v, u) = f(v) = 0,$$

so we can conclude $u_0 \in Z(\omega_0)$. We apply Lemma 4 to find a space $V_1 \subseteq Z(\omega_0)$ with $\dim V_1 \leq \ell^d$ and

$$\begin{aligned} \|u_0 - v_1\|_{L^2(\omega_0)} &\leq c_{\mathrm{apx}} \frac{\operatorname{diam}_\infty(\omega_0)}{\ell} \|\nabla u_0\|_{L^2(\omega_0)} \\ &\leq c_{\mathrm{apx}} \frac{\operatorname{diam}_\infty(\tau) + 2\delta}{\ell} \|\nabla u_0\|_{L^2(\omega_0)}. \end{aligned}$$

The admissibility assumption (5) implies $\operatorname{diam}_\infty(\tau) \leq 2\eta\delta$, and the estimate becomes

$$\|u_0 - v_1\|_{L^2(\omega_0)} \leq c_{\mathrm{apx}} \frac{2(\eta+1)\delta}{\ell} \|\nabla u_0\|_{L^2(\omega_0)}.$$

According to Lemma 2, the restriction $u_1 := (u_0|_{\omega_1} - v_1|_{\omega_1})$ of the error $u_0 - v_1$ is contained in $Z(\omega_1) \cap H^1(\omega_1)$ and the interior regularity estimate

$$\|\nabla u_1\|_{L^2(\omega_1)} \leq \frac{c_{\mathrm{reg}}}{\operatorname{dist}(\omega_1, \partial\omega_0)} \|u_0 - v_1\|_{L^2(\omega_0)}$$

$$\leq c_{\mathrm{reg}}\frac{p}{\delta}c_{\mathrm{apx}}\frac{2(\eta+1)\delta}{\ell}\|\nabla u_0\|_{L^2(\omega_0)} = c\frac{p}{\ell}\|\nabla u_0\|_{L^2(\omega_0)}$$

holds for the constant

$$c := 2c_{\mathrm{reg}}c_{\mathrm{apx}}(\eta+1).$$

We can apply the same argument to construct a space $V_2 \subseteq L^2(\omega_1)$ and a function $v_2 \in V_2$ approximating u_1, and proceed until we have found spaces $V_1,\ldots,V_p$ and functions $v_1 \in V_1,\ldots,v_p \in V_p$ with

$$\begin{aligned}\|u|_{\omega_i} - (v_1|_{\omega_i} + \ldots + v_i|_{\omega_i})\|_{L^2(\omega_i)} &\leq c_{\mathrm{apx}}\frac{2(\eta+1)\delta}{\ell}\left(c\frac{p}{\ell}\right)^{i-1}\|\nabla u_0\|_{L^2(\tau)}\\ &= \frac{\delta}{c_{\mathrm{reg}}p}\left(c\frac{p}{\ell}\right)^{i}\|\nabla u_0\|_{L^2(\tau)}\end{aligned}$$

and

$$\|\nabla(u|_{\omega_\ell} - (v_1|_{\omega_\ell} + \ldots + v_i|_{\omega_i}))\|_{L^2(\omega_\ell)} \leq \left(c\frac{p}{\ell}\right)^{i}\|\nabla u_0\|_{L^2(\tau)}$$

for all $i \in \{1,\ldots,p\}$, so due to $\tau \subseteq \omega_p$, the function

$$v := v_1|_\tau + \ldots + v_p|_\tau \in V := V_1|_\tau + \ldots + V_p|_\tau$$

is an approximation of $u_0|_\tau$ in the space $V := V_1|_\tau + \ldots + V_p|_\tau$ satisfying $\dim V \leq p\ell^d$ and the error estimates

$$\begin{aligned}\|\nabla(u|_\tau - v)\|_{L^2(\tau)} &\leq \left(c\frac{p}{\ell}\right)^{\ell}\|\nabla u_0\|_{L^2(\omega_0)},\\ \|u|_\tau - v\|_{H^1(\tau)} &= \left(\|u|_\tau - v\|^2_{L^2(\tau)} + \|\nabla(u|_\tau - v)|_\tau\|^2_{L^2(\tau)}\right)^{1/2}\\ &\leq \left(\frac{\delta^2}{c_{\mathrm{reg}}^2p^2}+1\right)^{1/2}\left(c\frac{p}{\ell}\right)^{p}\|\nabla u_0\|_{L^2(\omega_0)}\\ &\leq \left(\frac{\delta^2}{4^2 2^2}+1\right)^{1/2}\left(c\frac{p}{\ell}\right)^{p}\|\nabla u_0\|_{L^2(\omega_0)}\\ &\leq (\delta/8+1)\left(c\frac{p}{\ell}\right)^{p}\|\nabla u_0\|_{L^2(\omega_0)}.\end{aligned}$$

Since u is the solution of (2), we have

$$\|u\|^2_{H^1(\Omega)} \leq \frac{1}{C_\Omega\alpha}a(u,u) = \frac{1}{C_\Omega\alpha}f(u) \leq \frac{1}{C_\Omega\alpha}\|f\|_{H^{-1}(\Omega)}\|u\|_{H^1(\Omega)},$$

and this implies

$$\|u\|_{H^1(\Omega)} \leq \frac{1}{C_\Omega\alpha}\|f\|_{H^{-1}(\Omega)},$$

i.e.,

$$\|\nabla(u|_\tau - v)\|_{L^2(\tau)} \leq \frac{1}{C_\Omega\alpha}\left(c\frac{p}{\ell}\right)^{p}\|f\|_{H^{-1}(\Omega)},$$

$$\|u|_\tau - v\|_{H^1(\tau)} \leq \frac{1}{C_\Omega \alpha}(\delta/8+1)\left(c\frac{p}{\ell}\right)^p \|f\|_{H^{-1}(\Omega)}.$$

In order to get the estimates (4), (6) and (7), we have to choose ℓ appropriately. A simple approach is to let

$$\ell := \left\lceil \frac{cp^2}{q(p-1)} \right\rceil,$$

since this yields

$$c\frac{p}{\ell} \leq c\frac{pq(p-1)}{cp^2} = \frac{q(p-1)}{p} = q\left(1+\frac{1}{p-1}\right)^{-1},$$
$$\left(c\frac{p}{\ell}\right)^p \leq q^p\left(1+\frac{1}{p-1}\right)^{-p} \leq \frac{q^p}{e},$$

and the dimension of V can be bounded by

$$\begin{aligned}\ell &\leq \frac{cp^2}{q(p-1)} + 1 = \frac{cp(p-1)}{q(p-1)} + \frac{c(p-1)}{q(p-1)} + \frac{c}{q(p-1)} + 1 \\ &= \frac{c}{q}p + \frac{c}{q} + \frac{c}{q(p-1)} + 1 \leq \frac{c}{q}p + 2\frac{c}{q} + 1 \\ &\leq \frac{c}{q}p + \frac{c}{q}p + \frac{1}{2}p = \left(2\frac{c}{q}+\frac{1}{2}\right)p\end{aligned}$$

due to $p \geq 2$, so setting

$$C_{\mathrm{dim}} := \left(2\frac{c}{q}+\frac{1}{2}\right)^d, \qquad C_{\mathrm{apx}} := \frac{1}{C_\Omega \alpha}\frac{1}{e}$$

yields $\dim V \leq p\ell^d \leq C_{\mathrm{dim}} p^{d+1}$ and

$$\|\nabla(u|_\tau - v)\|_{L^2(\tau)} \leq C_{\mathrm{apx}} q^p \|f\|_{H^{-1}(\Omega)},$$
$$\|u|_\tau - v\|_{H^1(\tau)} \leq C_{\mathrm{apx}} q^p (\delta/8+1)\|f\|_{H^{-1}(\Omega)},$$

therefore the proof is complete. ■

This result is closely related to [2, Theorem 2.8], but it yields an H^1-norm estimate for the solution of the variational equation (2) using the H^{-1}-norm of the right-hand side functional instead of an L^2-norm estimate of Green's function. The main difference between both proofs is that the one given here exploits the fact that the original solution u already is L-harmonic in τ, therefore we can perform the approximation by Lemma 4 first, and follow it by the regularity estimate of Lemma 2 in order to get an H^1-estimate for the error. The proof of [2, Theorem 2.8], on the other hand, deals with Green's function, and this function is not globally in H^1, therefore the first step has to be the regularity estimate and the resulting error bound is given only for the L^2-norm.

Of course, we can use the same ordering of regularity estimates and approximation steps in Theorem 5 in order to get an estimate of the form

$$\|u|_\tau - v\|_{L^2(\tau)} \leq C_{\mathrm{apx}} q^p \|f\|_{H^{-1}(\Omega)}$$

instead of (6). Since the space V constructed in this way would differ from the one used in Theorem 5, we cannot simply combine both estimates in order to get an estimate for the full H^1-norm and have to rely on results of the type (7) instead.

4 Approximation of matrix blocks

Usually, strongly elliptic partial differential equations of the type (1) are treated numerically by a finite element method: a mesh $\mathcal{T}_h$ for the domain Ω is constructed, and nodal basis functions $(\varphi_i)_{i\in\mathcal{I}}$ are used to define a finite-dimensional space

$$V_h := \operatorname{span}\{\varphi_i \ : \ i \in \mathcal{I}\} \subseteq H_0^1(\Omega),$$

where $h \in \mathbb{R}_{>0}$ is the mesh width of the triangulation and $\mathcal{I}$ is the set of its interior nodes.

Using the standard Galerkin approach, an approximation $u_h \in V_h$ of u is represented in the form

$$u_h = \sum_{i\in\mathcal{I}} x_i \varphi_i$$

for the solution vector $x \in \mathbb{R}^{\mathcal{I}}$ of the linear system

$$Ax = b \tag{8}$$

given by the stiffness matrix $A \in \mathbb{R}^{\mathcal{I}\times\mathcal{I}}$ and the load vector $b \in \mathbb{R}^{\mathcal{I}}$ defined by

$$A_{ij} = a(\varphi_i, \varphi_j), \qquad b_i = f(\varphi_i), \qquad \text{for all } i, j \in \mathcal{I}. \tag{9}$$

The system (8) can be solved by several techniques, e.g., by fast direct solvers [16], multigrid iterations [13] or $\mathcal{H}$-matrix methods [14, 11, 5]. We focus on the latter approach: $\mathcal{H}$-matrices are data-sparse approximations of dense matrices. Many arithmetic operations like the matrix multiplication, inversion, or LU and Cholesky factorizations can be performed efficiently for $\mathcal{H}$-matrices. As far as our application is concerned, this means that we can compute an approximate $\mathcal{H}$-matrix inverse $\widetilde{S}$ of the stiffness matrix A and use it as a preconditioner for the system (8).

The problem of proving that the $\mathcal{H}$-matrix algorithms will yield a sufficiently accurate approximation of the inverse can be reduced to an existence result: since the adaptive arithmetic operations (cf. [10]) have a best-approximation property, we only have to show that an approximation of A^{-1} by an $\mathcal{H}$-matrix exists, because this already implies that the computed approximation $\widetilde{S}$ will be at least as good as this approximation.

This proof of existence can be accomplished using our main result stated in Theorem 5: a block $A^{-1}|_{t\times s}$ describes the mapping from a right-hand side vector b with support in

s to the restriction of the corresponding discrete solution to t. In order to apply our approximation result, we have to exploit the relationship between the inverse matrix A^{-1} and the inverse operator L^{-1}.

In [2], this problem is solved by applying L^2-orthogonal projections. Since these projections are non-local operators, additional approximation steps are required, which increase the rank, lead to sub-optimal error estimates, and make the overall proof quite complicated.

We propose a different approach: instead of a non-local L^2-projection, a Clément-type interpolation operator [8] can be used to map continuous functions into the discrete space. These operators are "sufficiently local" to provide us with improved error estimates and guarantee that the rank of the approximation will not deteriorate.

Let us recall the basic definitions and properties of Clément interpolation operators: for each $i \in \mathcal{I}$, we fix a functional $\lambda_i : L^2(\Omega) \to \mathbb{R}$ with $\operatorname{supp} \lambda_i \subseteq \operatorname{supp} \varphi_i$ satisfying the local projection property

$$\lambda_i(\varphi_j) = \delta_{ij} \qquad \text{for all } j \in \mathcal{I} \tag{10}$$

and the local stability property

$$\|\lambda_i(u)\varphi_i\|_{L^2(\Omega)} \leq C_{\mathrm{cs}} \|u\|_{L^2(\operatorname{supp}\varphi_i)} \qquad \text{for all } u \in L^2(\Omega) \tag{11}$$

for a constant $C_{\mathrm{cs}} \in \mathbb{R}_{>0}$ depending only on the shape-regularity of the mesh. Constructions of this kind can be found in [17, 3].

The interpolation operator is defined by

$$I_h : L^2(\Omega) \to V_h, \qquad u \mapsto \sum_{i \in V_h} \lambda_i(u)\varphi_i. \tag{12}$$

The local projection property (10) implies its global counterpart

$$I_h v_h = v_h \qquad \text{for all } v_h \in V_h, \tag{13}$$

and the local stability property (11) combined with the shape-regularity of the mesh yields the global stability property

$$\|I_h v\|_{L^2(\Omega)} \leq C_{\mathrm{cl}} \|v\|_{L^2(\Omega)} \qquad \text{for all } v \in L^2(\Omega) \tag{14}$$

with a constant $C_{\mathrm{cl}} \in \mathbb{R}_{>0}$ depending again only on the shape-regularity of the mesh.

Since the matrices we are dealing with are given with respect to the space $\mathbb{R}^{\mathcal{I}}$, not V_h, we need a way of switching between both spaces. This is handled by the standard basis isomorphism

$$\Phi : \mathbb{R}^{\mathcal{I}} \to V_h \subseteq H_0^1(\Omega), \qquad x \mapsto \sum_{i \in \mathcal{I}} x_i \varphi_i.$$

The interpolation operator I_h can be expressed by

$$I_h = \Phi\Lambda$$

if we define $\Lambda : L^2(\Omega) \to \mathbb{R}^{\mathcal{I}}$ by

$$(\Lambda v)_i := \lambda_i(v) \qquad \text{for all } i \in \mathcal{I}, v \in L^2(\Omega).$$

In order to construct the approximation of A^{-1} by using L^{-1} we have to turn a vector $b \in \mathbb{R}^{\mathcal{I}}$ into a functional, apply L^{-1}, and approximate the result again in V_h. The first step can be accomplished by using the adjoint of Λ: we define

$$\Lambda^* : \mathbb{R}^{\mathcal{I}} \to (L^2(\Omega))', \qquad b \mapsto (v \mapsto \langle b, \Lambda v \rangle_2).$$

This operator turns each vector in $\mathbb{R}^{\mathcal{I}}$ into a functional on $L^2(\Omega)$.

If the vector b is given by (9) for a right-hand side functional $f \in H^{-1}(\Omega)$, the projection property (10) implies

$$(\Lambda^* b)(\varphi_i) = b_i = f(\varphi_i) \qquad \text{for all } i \in \mathcal{I}, b \in \mathbb{R}^{\mathcal{I}},$$

therefore the functional $\Lambda^* b$ and the original right-hand side f of (1) yield the same Galerkin approximation u_h.

We have to prove that Λ^* is a bounded mapping with respect to the correct norms. We assume that the finite element mesh is shape-regular in the sense of [9, Definition 2.2], and a simple application of [9, Proposition 3.1] yields that there is a positive definite diagonal matrix $H \in \mathbb{R}^{\mathcal{I} \times \mathcal{I}}$ satisfying

$$C_{\mathrm{b1}} \|H^{d/2} x\|_2 \leq \|\Phi x\|_{L^2(\Omega)} \leq C_{\mathrm{b2}} \|H^{d/2} x\|_2 \qquad \text{for all } x \in \mathbb{R}^{\mathcal{I}},$$

where $C_{\mathrm{b1}}, C_{\mathrm{b2}} \in \mathbb{R}_{>0}$ are constants depending only on the shape-regularity of the mesh. Using this inequality, we can now prove the necessary properties of Λ^*:

Lemma 6 (Λ^* bounded and local) *Let $\alpha \in [0,1]$ and $b \in \mathbb{R}^{\mathcal{I}}$. We have*

$$\|\Lambda^* b\|_{H^{-1+\alpha}(\Omega)} \leq \frac{C_{\mathrm{cl}}}{C_{\mathrm{b1}}} \|H^{-d/2} b\|_2, \tag{15}$$

i.e., Λ^* is a continuous mapping from $\mathbb{R}^{\mathcal{I}}$ to $H^{-1+\alpha}(\Omega)$.

The mapping preserves locality, i.e., it satisfies

$$\operatorname{supp}(\Lambda^* b) \subset \bigcup \{\operatorname{supp} \varphi_i \ : \ i \in \mathcal{I}, b_i \neq 0\}. \tag{16}$$

Proof: Let $v \in H_0^1(\Omega)$. Let $y := \Lambda v$, $v_h := \Phi y = I_h v$ and $v_\perp = v - v_h$. Since the interpolation operator I_h is a projection, we have

$$I_h v_\perp = I_h(v - v_h) = I_h v - I_h I_h v = 0,$$

and since Φ is bijective, $0 = I_h v_\perp = \Phi \Lambda v_\perp$ implies $\Lambda v_\perp = 0$.

Due to the definition of Λ^*, we get

$$(\Lambda^* b)(v) = \langle b, \Lambda v \rangle_2 = \langle b, y \rangle_2 = \langle b, H^{-d/2} H^{d/2} y \rangle_2$$

$$
\begin{aligned}
&= \langle H^{-d/2} b, H^{d/2} y \rangle_2 \leq \|H^{-d/2} b\|_2 \|H^{d/2} y\|_2 \\
&\leq \frac{\|H^{-d/2} b\|_2}{C_{\mathrm{b1}}} \|\Phi y\|_2 = \frac{\|H^{-d/2} b\|_2}{C_{\mathrm{b1}}} \|v_h\|_{L^2(\Omega)} \\
&= \frac{\|H^{-d/2} b\|_2}{C_{\mathrm{b1}}} \|I_h v\|_{L^2(\Omega)} \leq \frac{C_{\mathrm{cl}}}{C_{\mathrm{b1}}} \|H^{-d/2} b\|_2 \|v\|_{L^2(\Omega)},
\end{aligned}
$$

and this implies (15).

The support of $f \in H^{-1}(\Omega)$ is defined as the smallest closed set such that $f(v) = 0$ holds for all $v \in H_0^1(\Omega)$ with $\operatorname{supp} v \cap \operatorname{supp} f = \emptyset$. In order to prove (16), we assume that

$$
\operatorname{supp} v \cap \operatorname{supp} \varphi_i = \emptyset \qquad \text{holds for all } i \in \mathcal{I} \text{ with } b_i \neq 0 \tag{17}
$$

and have to prove $(\Lambda^* b)(v) = 0$. Due to (17), we have $v|_{\operatorname{supp} \varphi_i} \equiv 0$ for $i \in \mathcal{I}$ with $b_i \neq 0$, and (17) implies

$$
\begin{gathered}
|y_i| = \frac{\|y_i \varphi_i\|_{L^2(\Omega)}}{\|\varphi_i\|_{L^2(\Omega)}} \leq \frac{C_{\mathrm{cs}}}{\|\varphi_i\|_{L^2(\Omega)}} \|v\|_{L^2(\operatorname{supp} \varphi_i)} = 0 \\
\text{for all } i \in \mathcal{I} \text{ with } b_i \neq 0,
\end{gathered}
$$

therefore $0 = \langle b, y \rangle_2 = (\Lambda^* b)(v)$, and this completes the proof. ∎

This result allows us to switch from the vector b corresponding to the discrete setting to the functional f of the variational setting. In the variational setting, we can apply Theorem 5 to construct the desired approximation of the solution, then we have to switch back to the discrete setting. Unfortunately, we cannot use the Galerkin projection to perform this last step, which would be the natural choice considering that we want to approximate u_h, since it is a global operator and the approximation result only holds for a subdomain. Therefore we have to rely on the Clément-type interpolation operator again, which has the desired locality property.

Using interpolation instead of the Galerkin projection leads to a second discrete approximation of L^{-1}, given by the matrix

$$
S = \Lambda L^{-1} \Lambda^* \in \mathbb{R}^{\mathcal{I} \times \mathcal{I}}.
$$

If we let $b \in \mathbb{R}^{\mathcal{I}}$, $f := \Lambda^* b$, $u := L^{-1} f$ and $\tilde{u}_h := I_h u$, we observe $\Phi S b = \tilde{u}_h$, i.e., S indeed provides us with the coefficients of the Clément-type approximation of the solution operator.

In general, we know that u_h will converge to u if the mesh width h tends to zero. Since I_h is L^2-stable, this means that due to

$$
\begin{aligned}
\|\tilde{u}_h - u\|_{L^2(\Omega)} &= \|I_h u - u\|_{L^2(\Omega)} = \|I_h(u - u_h) - (u - u_h)\|_{L^2(\Omega)} \\
&\leq (C_{\mathrm{cl}} + 1) \|u - u_h\|_{L^2(\Omega)},
\end{aligned} \tag{18}
$$

the interpolated solution $\tilde{u}_h$ will converge to the same limit.

If the equation (1) is $H^{1+\alpha}(\Omega)$-regular, it is possible to derive a refined error estimate for the matrices S and A^{-1}:

Lemma 7 (Clément vs. Galerkin) *Let $\alpha \in [0,1]$. We assume that for all functionals $f \in H^{-1+\alpha}(\Omega)$, the solution $u := L^{-1}f$ satisfies $u \in H_0^{1+\alpha}(\Omega)$ and*

$$\|u\|_{H^{1+\alpha}(\Omega)} \leq C_{\mathrm{rg}} \|f\|_{H^{-1+\alpha}(\Omega)}. \tag{19}$$

Then there is a constant $C_{\mathrm{cg}} \in \mathbb{R}_{>0}$ depending only on C_{rg}, C_{cl}, C_{b1}, and the shape-regularity of the mesh with

$$\|H^{d/2}(S - A^{-1})b\|_2 \leq C_{\mathrm{cg}} h^{2\alpha} \|H^{-d/2} b\|_2 \qquad \text{for all } b \in \mathbb{R}^{\mathcal{I}}.$$

Proof: Let $b \in \mathbb{R}^{\mathcal{I}}$, let $f := \Lambda^* b$ and $u := L^{-1}f$. Let $x := A^{-1}b$ and $u_h := \Phi x$. Let $\tilde{x} := Sb$ and $\tilde{u}_h := \Phi \tilde{x}$. By definition, we have $\tilde{u}_h = I_h u$.

Using the standard Aubin-Nitsche lemma yields

$$\|u - u_h\|_{L^2(\Omega)} \leq C_{\mathrm{an}} h^{2\alpha} \|f\|_{H^{-1+\alpha}(\Omega)},$$

with a constant C_{an} depending only on C_{rg} and the shape-regularity parameters of the mesh. Combining this estimate with (18) gives us

$$\begin{aligned}\|u_h - \tilde{u}_h\|_{L^2(\Omega)} &\leq \|u_h - u\|_{L^2(\Omega)} + \|u - \tilde{u}_h\|_{L^2(\Omega)} \\ &\leq C_{\mathrm{an}}(C_{\mathrm{cl}} + 2) h^{2\alpha} \|f\|_{H^{-1+\alpha}(\Omega)}.\end{aligned}$$

We observe

$$\begin{aligned}\|H^{d/2}(S - A^{-1})b\|_2 &\leq \frac{1}{C_{\mathrm{b1}}} \|\Phi S b - \Phi A^{-1} b\|_{L^2(\Omega)} \\ &= \frac{1}{C_{\mathrm{b1}}} \|\tilde{u}_h - u_h\|_{L^2(\Omega)} \leq \frac{C_{\mathrm{an}}(C_{\mathrm{cl}} + 2)}{C_{\mathrm{b1}}} h^{2\alpha} \|f\|_{H^{-1+\alpha}(\Omega)} \\ &\leq \frac{C_{\mathrm{an}}(C_{\mathrm{cl}} + 2)}{C_{\mathrm{b1}}} h^{2\alpha} \frac{C_{\mathrm{cl}}}{C_{\mathrm{b1}}} \|H^{-d/2} b\|_2\end{aligned}$$

and complete the proof be setting $C_{\mathrm{cg}} := C_{\mathrm{an}} C_{\mathrm{cl}}(C_{\mathrm{cl}} + 2)/C_{\mathrm{b1}}^2$. ∎

Due to this result, a good approximation of S on a sufficiently fine mesh is also a good approximation of A^{-1}, and a good approximation of S can be constructed by Theorem 5: we consider the approximation of $S|_{t \times s}$ for two sets $t, s \subseteq \mathcal{I}$ of indices. In order to apply our approximation result, we have to translate the index sets t and s into subdomains of $\mathbb{R}^d$. We do this by assuming that domains $B_t, D_s \subseteq \mathbb{R}^d$ satisfying

$$\operatorname{supp} \varphi_i \subseteq B_t, \qquad \operatorname{supp} \varphi_j \subseteq D_s \qquad \text{for all } i \in t, j \in s$$

are given. We require B_t to be convex, D_s is allowed to be non-convex.

Theorem 8 (Blockwise low-rank approximation) *Let $\eta \in \mathbb{R}_{>0}$ and $q \in (0,1)$. There are constants $C_{\mathrm{blk}}, C_{\mathrm{dim}} \in \mathbb{R}_{>0}$ depending only only on η, q, Ω and the shape regularity of the mesh such that for all $t, s \subseteq \mathcal{I}$ with*

$$\operatorname{diam}(B_t) \leq 2\eta \operatorname{dist}(B_t, D_s)$$

and all $p \in \mathbb{N}_{\geq 2}$ we can find a rank $k \in \mathbb{N}$ with $k \leq C_{\mathrm{dim}} p^{d+1}$ and matrices $X_{t,s} \in \mathbb{R}^{t \times k}$, $Y_{t,s} \in \mathbb{R}^{s \times k}$ with

$$\|H_t^{d/2}(S|_{t\times s} - X_{t,s}Y_{t,s}^\top)b\|_2 \leq C_{\mathrm{blk}} q^p \|H_s^{-d/2} b\|_2 \qquad \text{for all } b \in \mathbb{R}^s$$

for $H_t := H|_{t\times t}$, $H_s := H|_{s\times s}$, i.e., the submatrix of S corresponding to the block $t \times s$ can be approximated by a matrix of rank k.

Proof: Due to Theorem 5, there is a space $V \subseteq L^2(B_t)$ with $\dim V \leq C_{\mathrm{dim}} p^{d+1}$ and

$$\|u|_\tau - v\|_{H^1(\tau)} \leq C_{\mathrm{apx}}(\operatorname{dist}(B_t, D_s)/8 + 1) q^p \|f\|_{H^{-1}(\Omega)}$$

for $\tau := \bigcup\{\operatorname{supp}\varphi_i \ : \ i \in t\} \subseteq \Omega$, all $f \in H^{-1}(\Omega)$ with $\operatorname{supp} f \subseteq D_s$, $u := L^{-1}f$ and a function $v \in V$.

Let $b \in \mathbb{R}^s$. We extend b to a vector $\hat{b} \in \mathbb{R}^{\mathcal{I}}$ by

$$\hat{b}_i := \begin{cases} b_i & \text{if } i \in s, \\ 0 & \text{otherwise} \end{cases} \qquad \text{for all } i \in \mathcal{I}.$$

Due to Lemma 6, the functional $f := \Lambda^* \hat{b}$ satisfies

$$\operatorname{supp} f \subseteq D_s, \qquad \|f\|_{H^{-1}(\Omega)} \leq \frac{C_{\mathrm{cl}}}{C_{\mathrm{b1}}} \|H^{-d/2}\hat{b}\|_2 = \frac{C_{\mathrm{cl}}}{C_{\mathrm{b1}}} \|H_s^{-d/2} b\|_2. \tag{20}$$

Let $u := L^{-1}f$. Since v approximates u only locally, we need local variants of Λ and Φ:

$$\begin{aligned} \Lambda_t : L^2(\tau) &\to \mathbb{R}^t, & v &\mapsto (\lambda_i(v))_{i\in t}, \\ \Phi_t : \mathbb{R}^t &\to V_h, & y &\mapsto \sum_{i\in t} y_i \varphi_i. \end{aligned}$$

We let $\tilde{x} := \Lambda_t u$. According to the definition of S, we have $\tilde{x} = (S\hat{b})|_t = S|_{t\times s} b$.

Let us now turn our attention to the local approximation of u. We have already seen that we can find a function $v \in V$ with

$$\begin{aligned} \|u|_\tau - v\|_{H^1(\tau)} &\leq C_{\mathrm{apx}}(\operatorname{dist}(B_t, D_s)/8 + 1) q^p \|f\|_{H^{-1}(\Omega)} \\ &\leq C_{\mathrm{apx}}(\operatorname{diam}(\Omega)/8 + 1) q^p \|f\|_{H^{-1}(\Omega)}. \end{aligned} \tag{21}$$

We let $\tilde{y} := \Lambda_t v$ and observe that (11) implies

$$\|\varphi_i(\tilde{x}_i - \tilde{y}_i)\|_{L^2(\Omega)} \leq C_{\mathrm{cs}} \|u|_\tau - v\|_{L^2(\operatorname{supp}\varphi_i)} \qquad \text{for all } i \in \mathcal{I},$$

and due to the shape-regularity of the mesh this yields

$$\begin{aligned} \|H_t^{d/2}(\tilde{x} - \tilde{y})\|_{L^2(\tau)} &\leq \frac{1}{C_{\mathrm{b1}}} \|\Phi_t(\tilde{x} - \tilde{y})\|_{L^2(\tau)} \\ &= \frac{1}{C_{\mathrm{b1}}} \|\Phi_t \Lambda_t (u|_\tau - v)\|_{L^2(\tau)} \end{aligned}$$

$$= \frac{1}{C_{\mathrm{b1}}} \|I_h(u|_\tau - v)\|_{L^2(\tau)} \leq \frac{C_{\mathrm{cl}}}{C_{\mathrm{b1}}} \|u|_\tau - v\|_{L^2(\tau)}. \tag{22}$$

Now we can define

$$C_{\mathrm{blk}} := C_{\mathrm{apx}}(\operatorname{diam}(\Omega)/8 + 1) \frac{C_{\mathrm{cl}}^2}{C_{\mathrm{b1}}^2}$$

and combining (20), (21) and (22) yields

$$\|H_t^{d/2}(\tilde{x} - \tilde{y})\|_{L^2(\tau)} \leq C_{\mathrm{blk}} q^p \|H_s^{-d/2} b\|_2.$$

Using this result, we can now derive the low-rank approximation $X_{t,s} Y_{t,s}^\top$ of $S|_{t\times s}$: we introduce the space

$$Z_h := \{H_t^{d/2} \Lambda_t w \ : \ w \in V\}$$

and observe $k := \dim Z_h \leq \dim V \leq C_{\dim} p^{d+1}$ for the rank and $H_t^{d/2}\tilde{y} = H_t^{d/2}\Lambda_t v \in Z_h$. We fix an orthogonal basis of Z_h, i.e., a matrix $Q \in \mathbb{R}^{t\times k}$ with orthogonal columns and $\operatorname{range} Q = Z_h$.

We define $\tilde{z} := H_t^{-1/2} Q Q^\top H_t^{d/2} \tilde{x}$. Since Q is orthogonal, $QQ^\top$ is the orthogonal projection onto Z_h and we get

$$\langle H_t^{d/2}(\tilde{x} - \tilde{z}), w\rangle_2 = \langle H_t^{d/2}\tilde{x} - QQ^\top H_t^{d/2}\tilde{x}, QQ^\top w\rangle_2 = 0 \quad \text{for all } w \in Z_h,$$

i.e., $H_t^{d/2}\tilde{z}$ is the best approximation of $H_t^{d/2}\tilde{x}$ in the space Z_h. In particular, $H_t^{d/2}\tilde{z}$ is at least as good as $H_t^{d/2}\tilde{y}$, and we get

$$\|H_t^{d/2}(\tilde{x} - \tilde{z})\|_2 \leq \|H_t^{d/2}(\tilde{x} - \tilde{y})\|_2 \leq C_{\mathrm{blk}} q^p \|H_s^{-1/2} b\|_2.$$

We let

$$X_{t,s} := H_t^{-1/2} Q \in \mathbb{R}^{t\times k}, \qquad Y_{t,s} := S|_{t\times s}^\top H_t^{d/2} Q \in \mathbb{R}^{s\times k}$$

and conclude

$$\tilde{z} = H_t^{-1/2} Q Q^\top H_t^{d/2} \tilde{x} = (H_t^{-1/2} Q)(Q^\top H_t^{d/2} S|_{t\times s}) b = X_{t,s} Y_{t,s}^\top b,$$

which completes the proof. ■

5 Approximation by an $\mathcal{H}$-matrix

Using the blockwise approximation result of Theorem 8, we can now construct an $\mathcal{H}$-matrix approximation of S, and due to Lemma 7, this will also approximate the matrix A^{-1}.

We briefly recall the basic concepts of hierarchical matrices: they are based on a *cluster tree*, i.e., a tree $\mathcal{T}_{\mathcal{I}}$ satisfying

- $\operatorname{root}(\mathcal{T}_\mathcal{I}) = \mathcal{I}$,
- if $t \in \mathcal{T}_\mathcal{I}$ with $\operatorname{sons}(t) \neq \emptyset$, we have $t = \bigcup \{t' \ : \ t' \in \operatorname{sons}(t)\}$,
- if $t \in \mathcal{T}_\mathcal{I}$ with $\operatorname{sons}(t) = \emptyset$, we have $\#t \leq n_{\min}$.

The elements $t \in \mathcal{T}_\mathcal{I}$ of the cluster tree are called *clusters*. Based on the cluster tree, the index set $\mathcal{I} \times \mathcal{I}$ corresponding to a matrix $M \in \mathbb{R}^{\mathcal{I} \times \mathcal{I}}$ is split into a partition

$$P = \{t \times s \ : \ t, s \in \mathcal{T}_\mathcal{I}\}.$$

An *admissibility condition* is used to distinguish between admissible and inadmissible blocks in P: a block is admissible if we expect to be able to approximate if by low rank, and it is inadmissible otherwise.

Considering the requirements of Theorem 8, it is reasonable to introduce a family $(B_t)_{t \in \mathcal{T}_\mathcal{I}}$ of convex domains satisfying

$$\operatorname{supp} \varphi_i \subseteq B_t \qquad \text{for all } i \in t$$

and to consider a block $t \times s$ admissible if

$$\max\{\operatorname{diam}(B_t), \operatorname{diam}(B_s)\} \leq 2\eta \operatorname{dist}(B_t, B_s) \tag{23}$$

holds. Usually spheres or axis-parallel boxes are used for B_t, since they can be constructed by simple recursive algorithms. In order to keep the proofs simple, we also assume that $B_{t'} \subseteq B_t$ holds for all $t \in \mathcal{T}_\mathcal{I}$, $t' \in \operatorname{sons}(t)$. This is the case for all standard constructions.

We split P into admissible ("farfield") and inadmissible ("nearfield") blocks with respect to this condition:

$$P_{\text{far}} := \{t \times s \in P \ : \ (23) \text{ holds}\}, \qquad P_{\text{near}} := P \setminus P_{\text{far}}.$$

If a block $t \times s$ satisfies the condition (23), we can apply Theorem 8 to B_t and $D_s := B_s$ and get a rank $k \leq C_{\dim} p^{d+1}$ and matrices $X_{t,s} \in \mathbb{R}^{t \times k}$, $Y_{t,s} \in \mathbb{R}^{s \times k}$ with

$$\|H_t^{d/2}(S|_{t,s} - X_{t,s} Y_{t,s}^\top) b\|_2 \leq C_{\text{blk}} q^p \|H_s^{-1/2} b\|_2 \qquad \text{for all } b \in \mathbb{R}^s. \tag{24}$$

We define the $\mathcal{H}$-matrix approximation $\widetilde{S} \in \mathbb{R}^{\mathcal{I} \times \mathcal{I}}$ of S by

$$\widetilde{S}|_{t \times s} := \begin{cases} X_{t,s} Y_{t,s}^\top & \text{if } t \times s \in P_{\text{far}}, \\ S|_{t \times s} & \text{otherwise} \end{cases} \qquad \text{for all } t \times s \in P.$$

In order to derive an estimate for the error $\|S - \widetilde{S}\|_2$, we employ a simplified variant of the framework introduced in [10]: we define the level of a cluster $t \in \mathcal{T}_\mathcal{I}$ by

$$\operatorname{level}(t) := \begin{cases} \operatorname{level}(t^+) + 1 & \text{if there is a } t^+ \in \mathcal{T}_\mathcal{I} \text{ with } t \in \operatorname{sons}(t^+), \\ 0 & \text{otherwise, i.e., if } t = \operatorname{root}(\mathcal{T}_\mathcal{I}) \end{cases}$$

$$\text{for all } t \in \mathcal{T}_{\mathcal{I}}$$

and observe that the definition of the cluster tree implies that $\{t \in \mathcal{T}_{\mathcal{I}} \ : \ \operatorname{level}(t) = \ell\}$ is a disjoint partition of a subset of $\mathcal{I}$ for all $\ell \in \mathbb{N}_0$. In particular, we have

$$\sum_{\substack{t \in \mathcal{T}_{\mathcal{I}} \\ \operatorname{level}(t) = \ell}} \|x|_t\|_2^2 \leq \|x\|_2^2 \qquad \text{for all } x \in \mathbb{R}^{\mathcal{I}}, \ell \in \mathbb{N}_0. \tag{25}$$

The *block rows* and *block columns* for clusters are defined by

$$\begin{aligned} \operatorname{row}(t) &:= \{s \in \mathcal{T}_{\mathcal{I}} \ : \ t \times s \in P_{\text{far}}\}, \\ \operatorname{col}(t) &:= \{s \in \mathcal{T}_{\mathcal{I}} \ : \ s \times t \in P_{\text{far}}\} \qquad \text{for all } t \in \mathcal{T}_{\mathcal{I}}. \end{aligned}$$

The analysis of [10, 11] is based on the assumption that the cardinalities of the block rows and block columns are bounded, i.e., that there is a *sparsity constant* $C_{\text{sp}} \in \mathbb{N}$ satisfying

$$\begin{aligned} \#\operatorname{row}(t) &\leq C_{\text{sp}}, \\ \#\operatorname{col}(t) &\leq C_{\text{sp}} \qquad \text{for all } t \in \mathcal{T}_{\mathcal{I}}. \end{aligned}$$

In order to keep the presentation simple, we also assume that the partition P is *level-consistent*, i.e., that

$$\operatorname{level}(t) = \operatorname{level}(s) \qquad \text{holds for all } t \times s \in P.$$

A general construction ensuring that these conditions are fulfilled can be found in [11].

Lemma 9 (Spectral norm estimate) *Let $M \in \mathbb{R}^{\mathcal{I} \times \mathcal{I}}$, and let $x \in \mathbb{R}^{\mathcal{I}}$. Then we have*

$$\|Mx\|_2 \leq C_{\text{sp}} \left(\sum_{\ell=0}^{\infty} \max\{\|M|_{t \times s}\|_2 \ : \ t \times s \in P, \ \operatorname{level}(t) = \ell\} \right) \|x\|_2.$$

Proof: We introduce

$$\begin{aligned} \epsilon_{t,s} &:= \|M|_{t \times s}\|_2 \qquad && \text{for all } t \times s \in P, \\ \epsilon_\ell &:= \max\{\epsilon_{t,s} \ : \ t \times s \in P, \ \operatorname{level}(t) = \ell\} \qquad && \text{for all } \ell \in \mathbb{N}_0. \end{aligned}$$

For an arbitrary $y \in \mathbb{R}^{\mathcal{I}}$, the Cauchy-Schwarz inequality implies

$$\begin{aligned} \langle y, Mx \rangle_2 &= \sum_{t \times s \in P} \langle y|_t, M|_{t \times s} x|_s \rangle_2 \leq \sum_{t \times s \in P} \epsilon_{t,s} \|y|_t\|_2 \|x|_s\|_2 \\ &\leq \left(\sum_{t \times s \in P} \epsilon_{t,s} \|y|_t\|_2^2 \right)^{1/2} \left(\sum_{t \times s \in P} \epsilon_{t,s} \|x|_s\|_2^2 \right)^{1/2}, \end{aligned}$$

and (25) yields

$$\sum_{t\times s\in P} \epsilon_{t,s}\|x|_s\|_2^2 = \sum_{t\in\mathcal{T}_\mathcal{I}} \sum_{\substack{s\in\mathcal{T}_\mathcal{I}\\ t\times s\in P}} \epsilon_{t,s}\|x|_s\|_2^2 = \sum_{\ell=0}^{\infty} \sum_{\substack{t\in\mathcal{T}_\mathcal{I}\\ \operatorname{level}(t)=\ell}} \sum_{\substack{s\in\mathcal{T}_\mathcal{I}\\ t\times s\in P}} \epsilon_\ell\|x|_s\|_2^2$$
$$\leq C_{\mathrm{sp}} \sum_{\ell=0}^{\infty} \epsilon_\ell \sum_{\substack{t\in\mathcal{T}_\mathcal{I}\\ \operatorname{level}(t)=\ell}} \|x|_s\|_2^2 \leq C_{\mathrm{sp}} \sum_{\ell=0}^{\infty} \epsilon_\ell \|x\|_2^2,$$

and we conclude

$$\langle y, Mx\rangle_2 \leq C_{\mathrm{sp}} \sum_{\ell=0}^{\infty} \epsilon_\ell \|y\|_2\|x\|_2.$$

Applying this estimate to $y := Mx$ completes the proof. ∎

This result allows us to prove that $\widetilde{S}$ is indeed an $\mathcal{H}$-matrix approximation of the solution operator S:

Corollary 10 ($\mathcal{H}$-matrix approximation) *Let (24) hold for all admissible blocks $t\times s\in P$. Let $\varrho := \max\{\operatorname{level}(t) \ : \ t\in\mathcal{T}_\mathcal{I}\}$ be the depth of the cluster tree $\mathcal{T}_\mathcal{I}$. Then we have*

$$\|H^{d/2}(S-\widetilde{S})b\|_2 \leq C_{\mathrm{sp}}C_{\mathrm{blk}}(\varrho+1)q^p\|H^{-d/2}b\|_2 \qquad \text{for all } b\in\mathbb{R}^\mathcal{I}.$$

Proof: We introduce the matrix $E := H^{d/2}(S-\widetilde{S})H^{d/2}$. Since H is a diagonal matrix, we have

$$E|_{t\times s} = H_t^{d/2}(S|_{t\times s} - \widetilde{S}|_{t\times s})H_s^{1/2} \qquad \text{for all } t\times s\in P,$$

and (24) is equivalent with

$$\|E|_{t\times s}\hat{b}\|_2 = \|H_t^{d/2}(S-\widetilde{S})|_{t\times s}H_s^{1/2}\hat{b}\|_2 \leq C_{\mathrm{blk}}q^p\|\hat{b}\|_2$$
$$\text{for all } b\in\mathbb{R}^s, t\times s\in P_{\mathrm{far}},$$

therefore we can use Lemma 9 to get

$$\|H^{d/2}(S-\widetilde{S})H^{d/2}\hat{b}\|_2 = \|E\hat{b}\|_2 \leq C_{\mathrm{sp}}C_{\mathrm{blk}}(\varrho+1)q^p\|\hat{b}\|_2$$
$$\text{for all } \hat{b}\in\mathbb{R}^\mathcal{I}.$$

Substituting $\hat{b} = H^{-d/2}b$ yields the desired result. ∎

This means that S can be approximated by the $\mathcal{H}$-matrix $\widetilde{S}$ and that the accuracy of the approximation improves exponentially while the storage complexity grows only like a polynomial of order $d+1$.

6 Approximation by an $\mathcal{H}^2$-matrix

Let us now consider the approximation of S by an $\mathcal{H}^2$-matrix [15, 6]. $\mathcal{H}^2$-matrices combine concepts of $\mathcal{H}$-matrices and multilevel techniques: for each admissible block $t \times s \in P_{\text{far}}$, the corresponding submatrix is not only required to be of low rank, but its range and the range of its adjoint are required to be contained in special subspaces.

The subspaces are defined by *cluster bases*. Let $(V_t)_{t \in \mathcal{T}_\mathcal{I}}$ be a family of matrices satisfying $V_t \in \mathbb{R}^{t \times k}$ for a family of integers $(k_t)_{t \in \mathcal{T}_\mathcal{I}}$. It is called a *(nested) cluster basis* if for all $t \in \mathcal{T}_\mathcal{I}$ and $t' \in \operatorname{sons}(t)$ there is a matrix $E_{t'} \in \mathbb{R}^{k_{t'} \times k_t}$ with

$$V_t|_{t' \times k_t} = V_{t'} E_{t'},$$

i.e., if the matrices V_t for larger clusters can be expressed in terms of the matrices $V_{t'}$ of smaller clusters. In this context, the family $(k_t)_{t \in \mathcal{T}_\mathcal{I}}$ is called the *rank distribution* of the cluster basis.

The concept of nested cluster bases is similar to that of nested hierarchies of subspaces used in the analysis of multilevel techniques: the basis functions on coarser levels can be expressed by basis functions on finer levels, usually by a prolongation or interpolation operator.

Let $(V_t)_{t \in \mathcal{T}_\mathcal{I}}$ and $(W_s)_{s \in \mathcal{T}_\mathcal{I}}$ be cluster bases with the rank distributions $(k_t)_{t \in \mathcal{T}_\mathcal{I}}$ and $(\ell_s)_{s \in \mathcal{T}_\mathcal{I}}$. The matrix $M \in \mathbb{R}^{\mathcal{I} \times \mathcal{I}}$ is an $\mathcal{H}^2$-matrix with respect to these bases and the partition P if for each $t \times s \in P_{\text{far}}$ a *coupling matrix* $C_{t,s} \in \mathbb{R}^{k_t \times \ell_s}$ exists that satisfies

$$M|_{t \times s} = V_t C_{t,s} W_s^\top.$$

Since the rank of $M|_{t \times s}$ is bounded by $\min\{k_t, \ell_s\}$, each $\mathcal{H}^2$-matrix is also an $\mathcal{H}$-matrix, but the converse does not hold since V_t and W_s depend only on t or s, but not on both.

In order to prove that S can be approximated by an $\mathcal{H}^2$-matrix, we have to prove that suitable cluster bases $(V_t)_{t \in \mathcal{T}_\mathcal{I}}$ and $(W_s)_{s \in \mathcal{T}_\mathcal{I}}$ exist. We use the framework described in [4]: we define the *sets of descendants* by

$$\operatorname{sons}^*(t) := \begin{cases} \{t\} \cup \bigcup_{t' \in \operatorname{sons}(t)} \operatorname{sons}^*(t') & \text{if } \operatorname{sons}(t) \neq \emptyset, \\ \{t\} & \text{otherwise} \end{cases}$$
$$\text{for all } t \in \mathcal{T}_\mathcal{I}$$

and the *sets of predecessors* by

$$\operatorname{pred}(t) := \{t^+ \in \mathcal{T}_\mathcal{I} \ : \ t \in \operatorname{sons}^*(t^+)\} \qquad \text{for all } t \in \mathcal{T}_\mathcal{I}.$$

for each $t \in \mathcal{T}_\mathcal{I}$, we let

$$\operatorname{row}^*(t) := \bigcup \{\operatorname{row}(t^+) \ : \ t^+ \in \operatorname{pred}(t)\},$$
$$R_t := \bigcup \{s \ : \ s \in \operatorname{row}^*(t)\}.$$

The *total cluster basis* $(S_t)_{t \in \mathcal{T}_\mathcal{I}}$ is defined by

$$S_t := S|_{t \times R_t} \qquad \text{for all } t \in \mathcal{T}_\mathcal{I}.$$

Low-rank approximations of the matrices S_t give rise to suitable cluster bases:

Lemma 11 (Approximation of the total cluster basis) *We assume that for each $t \in \mathcal{T}_\mathcal{I}$ an accuracy $\epsilon_t \in \mathbb{R}_{>0}$, a rank $k_t \in \mathbb{N}$, and matrices $X_t \in \mathbb{R}^{t\times k_t}$, $Y_t \in \mathbb{R}^{R_t\times k_t}$ are given with*

$$\|S_t - X_t Y_t^\top\|_2 \leq \epsilon_t. \tag{26}$$

Then there is a cluster basis $(V_t)_{t\in\mathcal{T}_\mathcal{I}}$ with rank distribution $(k_t)_{t\in\mathcal{T}_\mathcal{I}}$ satisfying

$$\|S|_{t\times s} - V_t V_t^\top S|_{t\times s}\|_2^2 \leq \sum_{r\in\operatorname{sons}^*(t)} \epsilon_r^2 \qquad \text{for all } t\times s \in P_{\mathrm{far}}$$

and $V_t^\top V_t = I$ for all $t \in \mathcal{T}_\mathcal{I}$.

Proof: Due to [4, eq. (10)], restricting the global error estimate of [4, Theorem 3.13] to the submatrix $S|_{t\times s}$ gives us the desired estimate. ∎

We can use Theorem 8 to prove that the assumptions of Lemma 11 are satisfied.

Lemma 12 (Projection error) *Let $q \in (0,1)$. For all $p \in \mathbb{N}$, we can find cluster bases $(V_t)_{t\in\mathcal{T}_\mathcal{I}}$ and $(U_t)_{t\in\mathcal{T}_\mathcal{I}}$ with the rank distribution $(k_t)_{t\in\mathcal{T}_\mathcal{I}}$ such that*

$$\|H_t^{d/2}(S|_{t\times s} - V_t U_t^\top S|_{t\times s})H_s^{1/2}\|_2 \leq C_{\mathrm{blk}} q^p \sqrt{\#\operatorname{sons}^*(t)} \qquad \text{holds for all } t\times s \in P_{\mathrm{far}},$$

the ranks are bounded by $k_t \leq C_{\mathrm{dim}} p^{d+1}$ and

$$U_t^\top V_t = I \qquad \text{holds for all } t \in \mathcal{T}_\mathcal{I}.$$

Proof: Let $p \in \mathbb{N}$. We have to prove that we can find suitable low-rank approximations of the matrix S_t such that (26) holds. Let $t \in \mathcal{T}_\mathcal{I}$ and define

$$D_t := \bigcup \{B_s \ : \ s \in \operatorname{row}^*(t)\}.$$

For any $s \in \operatorname{row}^*(t)$, we can find $t^+ \in \operatorname{pred}(t)$ such that $t^+ \times s \in P_{\mathrm{far}}$, i.e., that the admissibility condition (23) holds for $B_{t^+} \times B_s$. Since $B_t \subseteq B_{t^+}$, we conclude

$$\operatorname{diam}(B_t) \leq \operatorname{diam}(B_{t^+}) \leq 2\eta \operatorname{dist}(B_{t^+}, B_s) \leq 2\eta \operatorname{dist}(B_t, B_s).$$

Since s is an arbitrary element of $\operatorname{row}^*(t)$, this means

$$\operatorname{diam}(B_t) \leq 2\eta \operatorname{dist}(B_t, D_t).$$

We apply Theorem 8 to find a rank $k_t \in \mathbb{N}$ with $k_t \leq C_{\mathrm{dim}} p^{d+1}$ and matrices $X_t \in \mathbb{R}^{t\times k_t}$ and $Y_t \in \mathbb{R}^{R_t\times k_t}$ with

$$\|H_t^{d/2}(S_t - X_t Y_t^\top)H_{R_t}^{1/2}\|_2 \leq C_{\mathrm{blk}} q^p. \tag{27}$$

In order to apply Lemma 11, we have to introduce the scaled matrices

$$\widehat{S} := H^{d/2} S H^{d/2}, \qquad \widehat{S}_t := H_t^{d/2} S_t H_{R_t}^{1/2},$$

$$\widehat{X}_t := H_t^{d/2} X_t, \qquad \widehat{Y}_t := H_{R_t}^{1/2} Y_t \qquad \text{for all } t \in \mathcal{T}_\mathcal{I}$$

and observe that (27) reads

$$\|\widehat{S}_t - \widehat{X}_t \widehat{Y}_t^\top\|_2 \leq C_{\mathrm{blk}} q^p.$$

We use Lemma 11 to get a cluster basis $(\widehat{V}_t)_{t \in \mathcal{T}_\mathcal{I}}$ with

$$\|\widehat{S}|_{t\times s} - \widehat{V}_t \widehat{V}_t^\top \widehat{S}|_{t\times s}\|_2^2 \leq C_{\mathrm{blk}}^2 \sum_{r \in \operatorname{sons}^*(t)} q^{2p} = C_{\mathrm{blk}}^2 q^{2p} \# \operatorname{sons}^*(t)$$
$$\text{for all } t \in \mathcal{T}_\mathcal{I}.$$

Now we can define the cluster bases $(V_t)_{t \in \mathcal{T}_\mathcal{I}}$ and $(U_t)_{t \in \mathcal{T}_\mathcal{I}}$ by

$$V_t := H_t^{-1/2} \widehat{V}_t, \qquad U_t := H_t^{d/2} \widehat{V}_t \qquad \text{for all } t \in \mathcal{T}_\mathcal{I}$$

and get

$$\begin{aligned}
&\|H_t^{d/2}(S|_{t\times s} - V_t U_t^\top S|_{t\times s}) H_s^{1/2}\|_2 \\
&\quad = \|(H_t^{d/2} S|_{t\times s} H_s^{1/2} - H_t^{d/2} V_t \widehat{V}_t^\top H_t^{d/2} S|_{t\times s} H_s^{1/2})\|_2 \\
&\quad = \|\widehat{S}|_{t\times s} - \widehat{V}_t \widehat{V}_t^\top \widehat{S}|_{t\times s}\|_2 \leq C_{\mathrm{blk}}^2 q^{2p} \# \operatorname{sons}^*(t) \\
&\qquad \text{for all } t \in \mathcal{T}_\mathcal{I}.
\end{aligned}$$

This is the desired result. ■

This error estimate suggests how to define the $\mathcal{H}^2$-matrix approximation of S: we define $\widetilde{S} \in \mathbb{R}^{\mathcal{I}\times\mathcal{I}}$ by projecting each admissible block, i.e., by setting

$$\widehat{S}|_{t\times s} := \begin{cases} V_t U_t^\top S|_{t\times s} U_s V_s^\top & \text{if } t \times s \in P_{\mathrm{far}}, \\ S|_{t\times s} & \text{otherwise} \end{cases} \qquad \text{for all } t \times s \in P_{\mathrm{far}}.$$

The coupling matrices are given by $C_{t,s} := U_t^\top S|_{t\times s} U_s$ for all $t \times s \in P_{\mathrm{far}}$, and we use $(V_t)_{t \in \mathcal{T}_\mathcal{I}}$ both as row and column cluster basis, since S is symmetric.

Corollary 13 ($\mathcal{H}^2$-matrix approximation) *Let q, p and the cluster bases $(V_t)_{t \in \mathcal{T}_\mathcal{I}}$, $(U_t)_{t \in \mathcal{T}_\mathcal{I}}$ with ranks $(k_t)_{t \in \mathcal{T}_\mathcal{I}}$ be as in Lemma 12. We assume that the cluster tree $\mathcal{T}_\mathcal{I}$ is not degenerate, i.e., that there are constants $C_{\mathrm{sn}} \in \mathbb{R}_{>0}$ and $\zeta \in (0,1)$ with*

$$\# \operatorname{sons}^*(t) \leq C_{\mathrm{sn}} \zeta^{\operatorname{level}(t)} \# \mathcal{T}_\mathcal{I} \qquad \textit{for all } t \in \mathcal{T}_\mathcal{I}.$$

Then we have

$$\|H^{d/2}(S - \widetilde{S}) b\|_2 \leq \frac{C_{\mathrm{sp}} C_{\mathrm{blk}} \sqrt{2 C_{\mathrm{sn}}}}{1 - \sqrt{\zeta}} q^p \sqrt{\# \mathcal{T}_\mathcal{I}} \|H^{-d/2} b\|_2$$
$$\textit{for all } b \in \mathbb{R}^\mathcal{I}.$$

Proof: Let $\widehat{S}$ and $(\widehat{V}_t)_{t\in\mathcal{T}_\mathcal{I}}$ be defined as in the proof of Lemma 12. Then we have

$$\begin{aligned}
\|H_t^{d/2}(S|_{t\times s}-\widetilde{S}|_{t\times s})H_s^{1/2}\|_2^2 &= \|\widehat{S}|_{t\times s}-\widehat{V}_t\widehat{V}_t^\top\widehat{S}|_{t\times s}\widehat{V}_s\widehat{V}_s^\top\|_2^2\\
&= \|\widehat{S}|_{t\times s}-\widehat{V}_t\widehat{V}_t^\top\widehat{S}|_{t\times s}+\widehat{V}_t\widehat{V}_t^\top(\widehat{S}|_{t\times s}-\widehat{S}|_{t\times s}\widehat{V}_s\widehat{V}_s^\top)\|_2^2\\
&= \|\widehat{S}|_{t\times s}-\widehat{V}_t\widehat{V}_t^\top\widehat{S}|_{t\times s}\|_2^2+\|\widehat{V}_t\widehat{V}_t^\top(\widehat{S}|_{t\times s}-\widehat{S}|_{t\times s}\widehat{V}_s\widehat{V}_s^\top)\|_2^2\\
&\leq \|\widehat{S}|_{t\times s}-\widehat{V}_t\widehat{V}_t^\top\widehat{S}|_{t\times s}\|_2^2+\|\widehat{S}|_{s\times t}-\widehat{V}_s\widehat{V}_s^\top\widehat{S}|_{s\times t}\|_2^2\\
&\leq C_{\mathrm{blk}}^2(q^{2p}\#\operatorname{sons}^*(t)+q^{2p}\#\operatorname{sons}^*(s))\\
&= C_{\mathrm{blk}}^2q^{2p}(\#\operatorname{sons}^*(t)+\#\operatorname{sons}(s))\\
&\leq C_{\mathrm{blk}}^2C_{\mathrm{sn}}q^{2p}(\zeta^{\operatorname{level}(t)}+\zeta^{\operatorname{level}(s)})\#\mathcal{T}_\mathcal{I}\\
&= 2C_{\mathrm{blk}}^2C_{\mathrm{sn}}q^{2p}\zeta^{\operatorname{level}(t)/2}\zeta^{\operatorname{level}(s)/2}\#\mathcal{T}_\mathcal{I}.
\end{aligned}$$

As in the proof of Corollary 10, we introduce the error matrix $E := H^{d/2}(S-\widetilde{S})H^{d/2}$. We have just proven

$$\|E|_{t\times s}\|_2 \leq C_{\mathrm{blk}}\sqrt{2C_{\mathrm{sn}}\#\mathcal{T}_\mathcal{I}}q^p\zeta^{\operatorname{level}(t)/2} \qquad \text{for all } t\times s\in P_{\mathrm{far}},$$

so we can apply Lemma 9 in order to get

$$\|E\|_2 \leq C_{\mathrm{sp}}C_{\mathrm{blk}}\sqrt{2C_{\mathrm{sn}}\#\mathcal{T}_\mathcal{I}}q^p\sum_{\ell=0}^{\infty}\zeta^{\ell/2} \leq \frac{C_{\mathrm{sp}}C_{\mathrm{blk}}\sqrt{2C_{\mathrm{sn}}}}{1-\sqrt{\zeta}}q^p\sqrt{\#\mathcal{T}_\mathcal{I}},$$

and the proof can be completed as in the case of Corollary 10. ■

While in the case of the $\mathcal{H}$-matrix estimate in Corollary 10 only the depth of the cluster tree appeared as an additional factor, the $\mathcal{H}^2$-matrix estimate in Corollary 13 involves a factor of $\sqrt{\#\mathcal{T}_\mathcal{I}} \sim \sqrt{\#\mathcal{I}}$. Fortunately, both factors can be compensated by increasing p: if the number of degrees of freedom grows, we would have to increase p anyway in order to keep pace with the improving discretization error, and the additional factors in the error estimates only mean that we have to increase p a little faster.

7 Numerical experiments

The usefulness of $\mathcal{H}$-matrix approximations of inverses and LU or Cholesky factorizations of stiffness matrices of elliptic partial differential equations has already been discussed in several papers (e.g., [11, 2, 12], to name just a few).

We can therefore focus on approximations using the more efficient $\mathcal{H}^2$-matrices and investigate whether they offer advantages compared to $\mathcal{H}$-matrix schemes.

Our experiments are carried out on the symmetric unit square $\Omega := [-1,1]^2$ with four different types of coeffient matrices. For basic tests, we use $C_C \equiv 1$ in (1), i.e., L will be the Laplace operator. Next, we investigate three problems with discontinuous

Table 1: Approximation of the inverse for Poisson's equation

	$\mathcal{H}$-matrix			$\mathcal{H}^2$-matrix		
n	Mem/n	Near/n	Error	Mem/n	Near/n	Error
1024	3.5	2.7	1.1_{-4}	2.6	2.0	1.1_{-4}
4096	5.6	3.2	3.5_{-4}	3.6	2.3	2.8_{-4}
16384	9.0	4.1	1.9_{-4}	4.7	2.5	1.6_{-4}
65536	12.6	4.3	2.4_{-4}	5.7	3.6	1.5_{-4}
262144	24.1	14.4	1.6_{-4}	6.5	3.7	1.3_{-4}

Table 2: Approximation of the inverse for the "quartered" coefficient C_Q

	$\mathcal{H}$-matrix			$\mathcal{H}^2$-matrix		
n	Mem/n	Near/n	Error	Mem/n	Near/n	Error
1024	2.6	2.0	1.5_{-4}	3.5	2.7	2.8_{-4}
4096	5.5	3.2	5.1_{-4}	3.7	2.3	3.8_{-4}
16384	8.9	4.1	3.2_{-4}	4.8	2.5	2.2_{-4}
65536	12.5	4.3	3.2_{-4}	5.8	3.6	2.0_{-4}
262144	23.9	14.4	2.1_{-4}	6.5	3.7	1.9_{-4}

coefficients: in the first problem, we separate the square Ω into four quarters and switch the coefficients between 1 and 100:

$$C_Q(x) := \begin{cases} 100 & \text{if } x \in [-1,0) \times [-1,0) \text{ or } x \in [0,1] \times [0,1], \\ 1 & \text{otherwise.} \end{cases}$$

In the second problem, we separate the lower and upper half of the square by a strip with high conductivity:

$$C_L(x) := \begin{cases} 100 & \text{if } x_2 \in [0, 1/16), \\ 1 & \text{otherwise.} \end{cases}$$

In the third problem, we introduce anisotropic coefficients in the lower half of the square:

$$C_A(x) := \begin{cases} I & \text{if } x_2 \in [-1, 0), \\ \operatorname{diag}(100, 1) & \text{otherwise.} \end{cases}$$

All of these coefficient functions satisfy the assumptions of our theory with the bounds $\alpha = 1$ and $\beta = 100$ for the spectrum.

The approximation of the inverse matrices A^{-1} depends on several parameters. Most important are $n_{\min} \in \mathbb{N}$ (cf. section 5), which determines how much of the matrix is stored as a standard dense matrix (the nearfield part), and the accuracy $\hat{\epsilon} \in \mathbb{R}_{>0}$ used during the adaptive computation process. In our experiments, we pick an $n_{\min}$

Table 3: Approximation of the inverse for the "line" coefficient C_L

	$\mathcal{H}$-matrix			$\mathcal{H}^2$-matrix		
n	Mem/n	Near/n	Error	Mem/n	Near/n	Error
1024	3.5	2.7	1.3_{-4}	2.6	2.0	9.3_{-5}
4096	5.5	3.2	3.2_{-4}	3.6	2.3	2.8_{-4}
16384	8.8	4.1	2.8_{-4}	4.7	2.5	1.6_{-4}
65536	12.5	4.3	2.3_{-4}	5.7	3.6	1.5_{-4}
262144	24.3	14.4	8.0_{-5}	6.7	3.7	6.4_{-5}

Table 4: Approximation of the inverse for the anisotropic coefficient C_A

	$\mathcal{H}$-matrix			$\mathcal{H}^2$-matrix		
n	Mem/n	Near/n	Error	Mem/n	Near/n	Error
1024	3.5	2.7	5.7_{-4}	2.6	2.0	2.5_{-4}
4096	5.6	3.2	1.2_{-3}	3.8	2.3	4.4_{-4}
16384	9.1	4.1	6.6_{-4}	5.0	2.5	4.4_{-4}
65536	13.3	4.3	3.3_{-4}	6.3	3.6	3.7_{-4}
262144	25.6	14.4	5.8_{-4}	7.2	3.7	4.7_{-4}

that ensures that the storage requirements of the near- and farfield parts of the matrix are roughly balanced. The accuracy $\hat{\epsilon}$ is chosen in such a way that the inversion error $\|I - \widetilde{S}A\|_2$ is less than 10^{-3}, which guarantees that the $\mathcal{H}$- or $\mathcal{H}^2$-matrix $\widetilde{S}$ is a very good preconditioner for the linear system.

The tables 1, 2, 3 and 4 list the storage requirements and inversion errors for $\mathcal{H}$- and $\mathcal{H}^2$-matrix approximations of A^{-1} with the coefficient functions C_C, C_Q, C_L and C_A introduced above. The columns "Mem/n" give the storage requirements per degree of freedom in KBytes, the columns "Near/n" give the nearfield part of the storage requirements, and the columns "Error" give an estimate for the inversion error $\|I - \widetilde{S}A\|_2$ computed by a power iteration.

We can see that the $\mathcal{H}^2$-matrix approximations always require less storage than their $\mathcal{H}$-matrix counterparts, although they reach a similar accuracy. For the $\mathcal{H}^2$-matrix approximation, the storage requirements seem to behave like $\mathcal{O}(n \log n)$, i.e., even better than the theoretical prediction of $\mathcal{O}(n \log^3 n)$.

References

[1] M. Bebendorf. Why approximate LU decompositions of finite element discretizations of elliptic operators can be computed with almost linear complexity. Technical Report 8/2005, Max Planck Institute for Mathematics in the Sciences, 2005. To appear in SIAM Journal of Numerical Analysis.

[2] M. Bebendorf and W. Hackbusch. Existence of $\mathcal{H}$-matrix approximants to the inverse FE-matrix of elliptic operators with L^∞-coefficients. *Numerische Mathematik*, 95:1–28, 2003.

[3] C. Bernardi and V. Girault. A local regularization operator for triangular and quadrilateral finite elements. *SIAM J. Num. Anal.*, 35(5):1893–1916, 1998.

[4] S. Börm. Data-sparse approximation of non-local operators by $\mathcal{H}^2$-matrices. *Linear Algebra and its Applications*, 422:380–403, 2007.

[5] S. Börm, L. Grasedyck, and W. Hackbusch. Hierarchical Matrices. Lecture Note 21 of the Max Planck Institute for Mathematics in the Sciences, 2003.

[6] S. Börm and W. Hackbusch. Data-sparse approximation by adaptive $\mathcal{H}^2$-matrices. *Computing*, 69:1–35, 2002.

[7] P. G. Ciarlet. *The finite element method for elliptic problems*. SIAM, 2002.

[8] P. Clément. Approximation by finite element functions using local regularization. *RAIRO Anal. Numér.*, 9:77–84, 1975.

[9] W. Dahmen, B. Faermann, I. G. Graham, W. Hackbusch, and S. A. Sauter. Inverse inequalities on non-quasiuniform meshes and applications to the mortar element method. *Math. Comp.*, 73:1107–1138, 2004.

[10] L. Grasedyck. *Theorie und Anwendungen Hierarchischer Matrizen*. Doctoral thesis, Universität Kiel, 2001.

[11] L. Grasedyck and W. Hackbusch. Construction and arithmetics of $\mathcal{H}$-matrices. *Computing*, 70:295–334, 2003.

[12] L. Grasedyck, R. Kriemann, and S. LeBorne. Parallel black box domain decomposition based $\mathcal{H}$-LU preconditioning. Technical Report 115, Max Planck Institute for Mathematics in the Sciences, Leipzig, 2005.

[13] W. Hackbusch. *Multi-Grid Methods and Applications*. Springer-Verlag Berlin, 1985.

[14] W. Hackbusch. A sparse matrix arithmetic based on $\mathcal{H}$-matrices. Part I: Introduction to $\mathcal{H}$-matrices. *Computing*, 62:89–108, 1999.

[15] W. Hackbusch, B. Khoromskij, and S. A. Sauter. On $\mathcal{H}^2$-matrices. In H. Bungartz, R. Hoppe, and C. Zenger, editors, *Lectures on Applied Mathematics*, pages 9–29. Springer-Verlag, Berlin, 2000.

[16] O. Schenk, K. Gärtner, W. Fichtner, and A. Stricker. PARDISO: A high-performance serial and parallel sparse linear solver in semiconductor device simulation. *Journal of Future Generation Computer Systems*, 18:69–78, 2001.

[17] L. R. Scott and S. Zhang. Finite element interpolation of nonsmooth functions satisfying boundary conditions. *Math. Comp.*, 54(190):483–493, 1990.