Max-Planck-Institut
für Mathematik
in den Naturwissenschaften
Leipzig

On the numerical solution of convection-dominated problems using hierarchical matrices

by

Mario Bebendorf

Preprint no.: 78 2005

On the numerical solution of convection-dominated problems using hierarchical matrices

M. Bebendorf
Fakultät für Mathematik und Informatik
Universität Leipzig
Augustusplatz 10/11
D-04109 Leipzig, Germany
bebendorf@math.uni-leipzig.de

September 2, 2005

Mathematics Subject Classification (2001): 35C20, 65F05, 65F50, 65N30
Keywords: singularly perturbed problems, preconditioners, hierarchical matrices

Abstract

The aim of this article is to shows that hierarchical matrices ($\mathcal{H}$-matrices) provide a means to efficiently precondition linear systems arising from the streamline diffusion finite-element method applied to convection-dominated problems. Approximate inverses and approximate LU decompositions can be computed with logarithmic-linear complexity in the standard $\mathcal{H}$-matrix format. Neither the complexity of the preconditioner nor the number of iterations will depend on the dominance. Although the established theory is only valid for irrotational convection, numerical experiments show that the same efficiency can be observed for general convection terms.

1 Introduction

We are concerned with the numerical solution of Dirichlet problems,

$$-\epsilon\Delta u + c\cdot\nabla u = f \quad \text{in } \Omega, \tag{1a}$$

$$u = g \quad \text{on } \partial\Omega, \tag{1b}$$

with $0 < \epsilon \ll |c|$ on bounded, simply connected Lipschitz domains $\Omega \subset \mathbb{R}^n$. Such kind of problems arise for instance from linearizing Navier-Stokes equations. Since the standard finite-element method on quasi-uniform grids is not uniformly stable with respect to ϵ, for the discretization of (1) the streamline diffusion finite-element method (SDFEM), see [26], or finite-difference methods, see [32], are used. Both methods provide ϵ-independent stability under reasonable assumptions.

In this article we concentrate on coefficient matrices $S \in \mathbb{R}^{N\times N}$ which have been generated from discretizing (1) by the SDFEM. Here and in the following, N denotes the number of degrees of freedom, which is assumed to be large. The Galerkin matrix S is sparse but has a bandwidth of order $N^{1-1/n}$, which due to fill-in leads to a complexity of direct methods of order $N^{3-2/n}$. Although recent implementations of the *LU* decomposition can hardly be beaten in two spatial dimensions, they are only suited for small problem sizes for three-dimensional problems. In the latter case iterative methods are usually more efficient.

Optimal complexity can be achieved by the *multigrid method* [20]. Since the convergence rate of multigrid methods depends on the coefficients of the operator and therefore on the parameter ϵ, in the last years much work has been done to develop robust multilevel methods. This is usually done by reordering techniques; see [28, 9].

For preconditioning Krylov subspace methods, whose convergence rates is determined by the distribution of eigenvalues of S, so-called sparse approximate inverse (SAI) preconditioners, see [7, 18, 12, 11] and [8] for an overview, have been introduced. In this approach the quantity $\|I - SC\|_F$ is usually minimized for matrices C having a given sparsity pattern. For a detailed survey on precondition techniques the reader is referred to [6].

The preconditioner presented in this article is related with the *incomplete LU factorization* (ILU) and its variants; see [31]. Similar to sparse inverses, the ILU is the *LU* decomposition computed on a given sparsity pattern, thereby avoiding fill-in. ILU is one of the best known and most often used preconditioning techniques. Although the ILU can be applied to any sparse coefficient matrix provided the factorization does not break down, it is well suited for M- and diagonally dominant matrices. ILU often significantly improves the convergence, but it does usually not lead to a bounded number of iterations of the solver. Instead of achieving an almost linear complexity by fixing the sparsity pattern of the factors L and U, we will compute low-rank approximations to suitable matrix blocks, i.e., the approximation of L and U will be done by so-called hierarchical matrices.

By the structure of hierarchical matrices ($\mathcal{H}$-matrices) introduced by Hackbusch et al. [22, 23] it is possible to treat fully populated matrices with logarithmic-linear complexity. In [2, 3] it was shown that the inverse of finite-element stiffness matrices of second order elliptic partial differential operators L with L^∞-coefficients can be approximated by $\mathcal{H}$-matrices. The proof relies on the relation

$$L^{-1}\varphi(x) = \int_\Omega G(x,y)\varphi(y)\,\mathrm{d}y \quad \text{for all } \varphi \in C_0^\infty(\Omega) \tag{2}$$

between the solution operator L^{-1} and the Green function G for L and Ω. This result shows that $\mathcal{H}$-matrices are robust in the sense that their efficiency does not depend on the smoothness and only slightly on the size of the coefficients. Low-precision approximants can therefore be used to precondition finite-element systems without any adaptation to the operator or to the geometry; see [4]. The results on the approximation of the inverse were used to prove the existence of $\mathcal{H}$-matrix approximants to the factors L and U of the *LU* decomposition; see [5]. Although approximations to both, the inverse and the *LU* decomposition, can be computed with almost linear complexity, the approximation of the factors L and U can be done significantly faster.

The proof in [3] on the approximation of inverses includes problems of type (1), but it does not account for the limiting case $\epsilon \to 0$ since the constants become unbounded as $\epsilon \to 0$. The aim of this article is to extend the existing theory to singularly perturbed problems. As a first step, we will consider the model problem (1) with the convection term c assumed to be irrotational. In this case it will be proved for SDFE stiffness matrices S that S^{-1} can be approximated by an $\mathcal{H}$-matrix which has logarithmic-linear complexity with ϵ-uniformly bounded constants. This result lays ground to the existence of approximate *LU* decompositions with almost linear complexity and hence to robust preconditioners for singularly perturbed problems. The algorithm for the computation of the preconditioner is exactly the same as it is used for computing preconditioners of general second-order elliptic operators. In [27] numerical experiments indicate that the $\mathcal{H}$-matrix format has to be adapted to dominating convection when approximating inverses of finite-difference matrices even for constant convection. We remark that our theory and also the numerical experiments from this article show that the SDFE discretization allows to employ the usual $\mathcal{H}$-matrix format, which does not depend on the coefficients of the operator.

The structure of the rest of this article is as follows. In Section 2 a brief review of the structure of $\mathcal{H}$-matrices together with the existence results for the approximation of the inverse and the LU decomposition will be given. Section 3 contains the existence theory of degenerate kernel approximants of the Green function for operators L of type (1), i.e.,

$$G(x,y) \approx \sum_{i=1}^{k} u_i(x) v_i(y), \quad x \in D_1 \text{ and } y \in D_2,$$

on an appropriate pair of domains (D_1, D_2). In Section 4 this result is then employed using (2) to show that the discrete inverse of L can be approximated by $\mathcal{H}$-matrices, which in turn leads to the existence of $\mathcal{H}$-matrix approximants to the inverse stiffness matrix. The numerical experiments from Section 5 will support our theory. It will be see that the hierarchical LU decomposition can be computed with almost linear complexity independently of ϵ. In addition, we employ this approximate LU decomposition to precondition GMRES applied to SDFE discretizations of (1). From these experiments a problem-independent (especially ϵ-independent) convergence rate can be observed for both, irrotational and rotational convection terms.

2 Hierarchical matrices

This section gives a brief overview over the structure of $\mathcal{H}$-matrices introduced by Hackbusch et al. [22, 23]. Roughly speaking, $\mathcal{H}$-matrices are matrices which are blockwise low-rank on partitions P stemming from recursive subdivision of the set of matrix indices. They provide a means to handle fully populated matrices with almost linear complexity. The combination of low-rank matrices and recursive subdivisions was also used in the mosaic-skeleton method [33].

Usually, matrices cannot be represented by $\mathcal{H}$-matrices without approximation. In order to exploit their efficiency, it is crucial to know whether a given matrix A can be approximated, i.e., for each block $b = t \times s \in P$

$$A_b \approx UV^T, \quad U \in \mathbb{R}^{t\times k},\, V \in \mathbb{R}^{s\times k},$$

and, what is even more important, that the required blockwise rank k is small compared with $|t|$ and $|s|$. A logarithmic dependence on the approximation accuracy and hence on the number of degrees of freedom is desirable. Obviously, by A_b we denote the subblock in the intersection of the rows t and columns s of A.

From Example 2.1 it will be seen that the stiffness matrix $S \in \mathbb{R}^{N\times N}$ with

$$s_{ij} = a(\varphi_j, \psi_i), \quad i,j \in I := \{1,\dots,N\}, \tag{3}$$

of operators L of the general type (6) possesses this property. In fact, S can be represented in the $\mathcal{H}$-matrix format without approximation. In the last expression, φ_j and ψ_i denote ansatz and trial functions and a is a bilinear form. While S is sparse, its inverse and the factors of its

LU decomposition are fully populated. In Section 2.3 we will review the theory on the existence of $\mathcal{H}$-matrix approximants for the latter matrices.

In contrast to other efficient methods like wavelet techniques [10, 13, 14], fast multipole and panel clustering, see [17], [24] and the references therein, $\mathcal{H}$-matrices concentrate on the matrix level. They are purely algebraic in the sense that once the $\mathcal{H}$-matrix approximant is built, no further information about the underlying problem is needed.

2.1 Admissibility condition

In order to be able to approximate each block b of a given matrix A by a low-rank matrix, b usually has to satisfy a certain condition. This so-called *admissibility condition* is the criterion for choosing whether a block b belongs to P. In the field of elliptic partial differential equations the following condition on $b = t \times s$ has proved useful:

$$\min\{\operatorname{diam} X_t, \operatorname{diam} Y_s\} \leq \eta \operatorname{dist}(X_t, Y_s), \tag{4}$$

where $\eta > 0$ is a given real number from the interval $[0.5, 1.5]$ and

$$X_t := \bigcup_{j \in t} X_j, \quad Y_s := \bigcup_{i \in s} Y_i$$

is the union of the supports of the basis functions ψ_i, $i \in t$, and φ_j, $j \in s$. We will see that under quite general assumptions this condition allows to approximate the Green function of L by a degenerate kernel, i.e., there are functions u_i, v_i, $i = 1, \ldots, k$, so that

$$G(x, y) \approx \sum_{i=1}^{k} u_i(x) v_i(y) \qquad \text{in } X_t \times Y_s, \tag{5}$$

where k depends only logarithmically on N. The degenerate approximation of G on $X_t \times Y_s$ will finally lead to a low-rank approximation of the block b. The construction of P can be done automatically from a given polyhedrization of Ω with almost linear complexity; see [1, 16] for two similar concepts.

The set of $\mathcal{H}$-matrices for a partition P with blockwise rank k is defined as

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

Note that $\mathcal{H}(P, k)$ is not a linear space, since the sum of two rank-k matrices exceeds rank k in general.

Example 2.1. *The stiffness matrix S of the differential operator L from* (6) *is in $\mathcal{H}(P, n_{\min})$, where $n_{\min}$ denotes the minimal blocksize. If $b \in P$ satisfies* (4)*, then the supports of the basis functions are pairwise disjoint. Hence, the matrix entries in this block vanish. In the remaining case b does not satisfy* (4)*. Then the size of one of the clusters is less than or equal to $n_{\min}$. In both cases the rank of S_b does not exceed $n_{\min}$.*

2.2 Storage and accelerated matrix operations

The cost of multiplying an $\mathcal{H}$-matrix $A \in \mathcal{H}(P, k)$ and its transposed A^T by a vector $x \in \mathbb{R}^N$ is inherited from the blockwise matrix-vector multiplication

$$Ax = \sum_{t \times s \in P} A_{t \times s} x_s \quad \text{and} \quad A^T x = \sum_{t \times s \in P} (A_{t \times s})^T x_t.$$

Since each block $t \times s$ has the representation $A_{t\times s} = UV^T$, $U \in \mathbb{R}^{t\times k}$, $V \in \mathbb{R}^{s\times k}$, $\mathcal{O}(k(|t| + |s|))$ units of memory are needed to store $A_{t\times s}$ and the matrix-vector products

$$A_{t\times s}x_s = UV^T x_s \quad \text{and} \quad (A_{t\times s})^T x_t = VU^T x_t$$

can be done with $\mathcal{O}(k(|t| + |s|))$ operations. Exploiting the hierarchical structure of A, it can therefore be shown that both storing A and multiplying A and A^T by a vector has $\mathcal{O}(\eta^{-n}kN \log N)$ complexity. For a rigorous analysis the reader is referred to [1]. Therefore, $\mathcal{H}$-matrices are well suited for iterative schemes such as Krylov subspace methods. In addition to the exact matrix-vector product also approximate versions of the usual matrix operations such as addition, multiplication and inversion can be defined using approximate block operations. These algorithms can be shown to have almost linear complexity; cf. [22, 23, 16].

2.3 Approximation of FE inverses

In [2, 3] we have considered general second order elliptic operators

$$Lu = -\text{div}\,[A\nabla u + bu] + c \cdot \nabla u + du \tag{6}$$

on bounded Lipschitz domains $\Omega \subset \mathbb{R}^n$. Let $\kappa \in \mathbb{R}$ denote an upper bound for the ratio of the largest and the smallest eigenvalue of $A(x)$, $x \in \Omega$. In [3, Thm. 3.5] it was shown that the Green function of such operators L can be approximated on domains which are far enough away from each other.

Theorem 2.2. *Let $D_1, D_2 \subset \mathbb{R}^n$ satisfy $\text{dist}(D_1, D_2) \geq \eta\,\text{diam}\,D_2$. Then for any $\delta > 0$ there are functions u_i, v_i with $Lv_i = 0$, $i = 1, \ldots, k$, in D_2 such that*

$$G_k(x, y) := \sum_{i=1}^{k} u_i(x)v_i(y)$$

satisfies

$$\|G(x, \cdot) - G_k(x, \cdot)\|_{L^2(D_2)} \leq \delta \|G(x, \cdot)\|_{L^2(\hat{D}_2)} \quad \text{for all } x \in D_1,$$

where $\hat{D}_2 := \{y \in \Omega : 2\eta\text{dist}(y, D_2) < \text{diam}\,D_2\}$. For the degree of degeneracy k it holds that $k \leq c_L^n |\log \delta|^{n+1} + |\log \delta|$, where

$$c_L := 2c_A e(1 + \eta) \left(\left(4\sqrt{\kappa} + \frac{\delta}{\lambda} \||b| + |c|\|_\infty \right)^2 + 8\frac{\delta}{\lambda}\||b|\|_\infty + 2\frac{\delta^2}{\lambda}\|d\|_\infty \right)^{1/2} \tag{7}$$

depends only on the size of the coefficients of L. If $d \geq 0$, then d does not appear in (7).

This result was used in [3] to show that the inverse FE stiffness matrix S can be approximated by an $\mathcal{H}$-matrix with blockwise rank at most k.

Based on the previous result, in [5] a similar result for the approximation of the factors L and U of the LU decomposition of S has been proved. In the following theorem ρ_N denotes the *growth factor*, cf. [25], which plays a central role for the stability of the LU decomposition.

Theorem 2.3. *There are lower and upper triangular matrices $L_\mathcal{H}, U_\mathcal{H} \in \mathcal{H}(P, k)$ with*

$$k \sim (\log N)^2 \left[|\log \delta| + (\log N)^2 + (\log N)(\log \rho_N \text{cond}_2 A) \right]^{n+1}$$

such that

$$\|A - L_\mathcal{H} U_\mathcal{H}\|_2 \leq \delta.$$

The asymptotic complexity of computing approximations of the inverse and of the factors of an LU decomposition is inherited from the approximate matrix multiplication. Hence, computing these approximations in $\mathcal{H}(P,k)$ can be done with $\mathcal{O}(k^2 N \log^2 k)$ arithmetical operations. We remark that the computation of an approximation to the factors of the LU decomposition is significantly faster than approximating the inverse. Since forward/backward substitution for $\mathcal{H}$-matrices can be done as fast the hierarchical matrix-vector multiplication, see [5], the LU decomposition should be used instead of the inverse. Hence, we will only use $\mathcal{H}$-matrix inverses for theoretical purposes.

3 Application to convection-dominated problems

As a model problem we consider the convection-dominated linear system

$$-\epsilon \Delta u + c \cdot \nabla u = f \quad \text{in } \Omega \tag{8a}$$

$$u = 0 \quad \text{on } \partial\Omega, \tag{8b}$$

where $0 < \epsilon \ll 1$ and $c \in (H^1(\Omega))^n$, $|c(x)| = 1$ for all x from the simply connected Lipschitz domain $\Omega \subset \mathbb{R}^n$. Since we are particularly interested in the limiting case $\epsilon \to 0$, we may assume that $2\epsilon \operatorname{div} c \leq 1$.

The operator $L : H_0^1(\Omega) \to H^{-1}(\Omega)$ defined by

$$Lu = -\epsilon \Delta u + c \cdot \nabla u$$

is an invertible second-order partial differential operator of type (6). If we try to apply the results from Theorem 2.2 to convection-dominated problems, the constant c_L from (7) will read

$$c_L = 2c_A e(1+\eta)\left(4 + \frac{\delta}{\epsilon}\right),$$

which is unbounded for the critical limit $\epsilon \to 0$. Since this constant enters the rank estimate, our existing theory does not show boundedness of the rank with respect to ϵ.

The aim of this article is to present another approach which will guarantee the desired boundedness. For this purpose we assume that the convection is an irrotational vector field, i.e., $\operatorname{curl} c = 0$. The following lemma shows that after appropriate transformation this problem can be looked at as a diffusion-reaction problem.

Lemma 3.1. *Assume that* $\operatorname{curl} c = 0$. *Then there is* $\phi_\epsilon : \Omega \to \mathbb{R}$ *such that*

$$-\epsilon \Delta u + c \cdot \nabla u = \epsilon e^{\phi_\epsilon}\left[-\Delta v + \frac{1}{2\epsilon}\left(\frac{1}{2\epsilon} - \operatorname{div} c\right) v\right],$$

where $v = e^{-\phi_\epsilon} u$.

Proof. Since $\operatorname{curl} c = 0$ and since Ω is simply connected, a potential ϕ_ϵ exists such that

$$\nabla \phi_\epsilon = \frac{1}{2\epsilon} c \quad \text{for all } x \in \Omega. \tag{9}$$

We first observe that $\nabla(e^{-\phi_\epsilon} u) = e^{-\phi_\epsilon}(\nabla u - u \nabla \phi_\epsilon)$. From

$$\operatorname{div} e^{-\phi_\epsilon} \nabla u = e^{-\phi_\epsilon}\left(\Delta u - \nabla \phi_\epsilon \cdot \nabla u\right)$$

and

$$\operatorname{div} e^{-\phi_\epsilon} u \nabla \phi_\epsilon = e^{-\phi_\epsilon} \left(\nabla u \cdot \nabla \phi_\epsilon - u |\nabla \phi_\epsilon|^2 + u \Delta \phi_\epsilon \right)$$

we obtain $\Delta v = \operatorname{div} \nabla (e^{-\phi_\epsilon} u) = e^{-\phi_\epsilon} \left[\Delta u - 2 \nabla \phi_\epsilon \cdot \nabla u + |\nabla \phi_\epsilon|^2 u - u \Delta \phi_\epsilon \right]$. Using (9), we are led to

$$e^{\phi_\epsilon} \Delta v - \frac{1}{2\epsilon} (\frac{1}{2\epsilon} - \operatorname{div} c) u = \Delta u - \epsilon^{-1} c \cdot \nabla u,$$

which proves the assertion. □

Hence, the last lemma proves the representation

$$L = \epsilon e^{\phi_\epsilon} \hat{L} e^{-\phi_\epsilon}$$

if we define

$$\hat{L} = -\Delta + \frac{1}{2\epsilon} \left(\frac{1}{2\epsilon} - \operatorname{div} c \right).$$

3.1 Degenerate approximation of the Green function

It is shown in [19] that in the case $n \geq 3$ a Green function $\hat{G} : \Omega \times \Omega \to \mathbb{R} \cup \{\infty\}$ for $\hat{L}$ and Ω exists with the properties

$$\hat{G}(x, \cdot) \in H^1(\Omega \setminus B_r(x)) \cap W_0^{1,1}(\Omega) \text{ for all } x \in \Omega \text{ and all } r > 0, \tag{10a}$$

$$\hat{a}(\hat{G}(x, \cdot), \varphi) = \varphi(x) \text{ for all } \varphi \in C_0^\infty(\Omega) \text{ and } x \in \Omega, \tag{10b}$$

where $B_r(x)$ is the open ball centered at x with radius r and

$$\hat{a}(u, v) = \int_\Omega \nabla v \cdot \nabla u \, \mathrm{d}x + \frac{1}{2\epsilon} \left(\frac{1}{2\epsilon} - \operatorname{div} c \right) uv \tag{11}$$

denotes the bilinear form associated with $\hat{L}$. Using $\hat{G}$ we define

$$G(x, y) = \epsilon^{-1} e^{\phi_\epsilon(y)} \hat{G}(x, y) e^{-\phi_\epsilon(x)}.$$

Then $G(x, \cdot) \in H^1(\Omega \setminus B_r(x)) \cap W_0^{1,1}(\Omega)$ for all $x \in \Omega$ and all $r > 0$. Furthermore,

$$\begin{aligned} (L_y G(x, y), \varphi(y))_{L^2} &= (e^{\phi_\epsilon(y)} \hat{L}_y e^{-\phi_\epsilon(y)} e^{\phi_\epsilon(y)} \hat{G}(x, y) e^{-\phi_\epsilon(x)}, \varphi(y))_{L^2} \\ &= (e^{\phi_\epsilon(y) - \phi_\epsilon(x)} \hat{L}_y \hat{G}(x, y), \varphi(y))_{L^2} \\ &= (\hat{L}_y \hat{G}(x, y), e^{\phi_\epsilon(y) - \phi_\epsilon(x)} \varphi(y))_{L^2} = \varphi(x) \end{aligned}$$

for all $\varphi \in C_0^\infty(\Omega)$. Therefore, G is a Green function for L and Ω. The following lemma is stated for later use.

Lemma 3.2. *For the Green function G for L and Ω it holds that $G \geq 0$.*

Proof. Since $G(x, y) = \epsilon^{-1} e^{\phi_\epsilon(y)} \hat{G}(x, y) e^{-\phi_\epsilon(x)}$, it is sufficient to show that $\hat{G}(x, y) \geq 0$. For this purpose let $y \in \Omega$ be fixed. We employ a technique that was also used in [19]. The bilinear form

$$\hat{a}(u, v) = \int_\Omega \nabla u \cdot \nabla v + \frac{1}{2\epsilon} \left(\frac{1}{2\epsilon} - \operatorname{div} c \right) uv$$

associated with the diffusion-reaction problem is continuous, symmetric and positive definite on $H_0^1(\Omega) \times H_0^1(\Omega)$. By the Lax-Milgram theorem for each $\rho > 0$ there is $\hat{G}_\rho \in H_0^1(\Omega)$, such that

$$\hat{a}(\hat{G}_\rho, \varphi) = \frac{1}{\operatorname{vol} B_\rho(y)} \int_{B_\rho(y)} \varphi \quad \text{for all } \varphi \in H_0^1(\Omega).$$

Note that $\hat{G}_\rho \to G$ for $\rho \to 0$. In particular, we obtain

$$\hat{a}(\hat{G}_\rho, \hat{G}_\rho) = \frac{1}{\operatorname{vol} B_\rho(y)} \int_{B_\rho(y)} \hat{G}_\rho \leq \frac{1}{\operatorname{vol} B_\rho(y)} \int_{B_\rho(y)} |\hat{G}_\rho| = \hat{a}(\hat{G}_\rho, |\hat{G}_\rho|).$$

This defines $r \leq 1$ such that $\hat{a}(\hat{G}_\rho, \hat{G}_\rho) = \hat{a}(\hat{G}_\rho, r|\hat{G}_\rho|)$. From

$$\hat{a}(r|\hat{G}_\rho|, r|\hat{G}_\rho|) = r^2 \hat{a}(\hat{G}_\rho, \hat{G}_\rho) \leq \hat{a}(\hat{G}_\rho, r|\hat{G}_\rho|)$$

we obtain

$$\hat{a}(r|\hat{G}_\rho| - \hat{G}_\rho, r|\hat{G}_\rho| - \hat{G}_\rho) = \hat{a}(r|\hat{G}_\rho|, r|\hat{G}_\rho|) - 2\hat{a}(\hat{G}_\rho, r|\hat{G}_\rho|) + \hat{a}(\hat{G}_\rho, \hat{G}_\rho) \leq 0.$$

Hence, $\hat{G}_\rho = r|\hat{G}_\rho| \geq 0$. The assertion follows in the limit $\rho \to 0$. □

Our aim is to define a degenerate approximant for G by a degenerate approximant for $\hat{G}$. From Theorem 2.2 we know that on an admissible pair (D_1, D_2) of domains for any $\delta > 0$ there are functions $\hat{u}_i$, $\hat{v}_i$ with $\hat{L}\hat{v}_i = 0$, $i = 1, \ldots, k$, in D_2 such that $\hat{G}_k(x,y) := \sum_{i=1}^k \hat{u}_i(x)\hat{v}_i(y)$ satisfies

$$\|\hat{G}(x,\cdot) - \hat{G}_k(x,\cdot)\|_{L^2(D_2)} \leq \delta \|\hat{G}(x,\cdot)\|_{L^2(\hat{D}_2)} \quad \text{for all } x \in D_1, \tag{12}$$

where $\hat{D}_2 := \{y \in \Omega : 2\eta \operatorname{dist}(y, D_2) < \operatorname{diam} D_2\}$. Since we may assume that $2\epsilon \operatorname{div} c \leq 1$, the reaction coefficient

$$\frac{1}{2\epsilon}\left(\frac{1}{2\epsilon} - \operatorname{div} c\right)$$

is non-negative and will therefore not appear in the bound on the degeneracy k. Hence, we have that $k \leq c_{\hat{L}}^n |\log \delta|^{n+1} + |\log \delta|$ with

$$c_{\hat{L}} = 8 c_A e(1+\eta)$$

is bounded independently of ϵ.

Instead of the L^2 estimate (12) we rather need a pointwise estimate. Exploiting the elliptic nature of the problem, this is derived in the following lemma.

Lemma 3.3. *Assume that* $\operatorname{dist}(D_1, D_2) \geq \eta \operatorname{diam} D_2$. *Then for any* $\delta > 0$ *there are functions* $\hat{u}_i$, $\hat{v}_i$, $i = 1, \ldots, k$, *such that* $\hat{G}_k(x,y) := \sum_{i=1}^k \hat{u}_i(x)\hat{v}_i(y)$ *satisfies*

$$|\hat{G}(x,y) - \hat{G}_k(x,y)| \leq \delta |\hat{G}(x,y)| \quad \textit{for all } x \in D_1 \textit{ and all } y \in D_2, \tag{13}$$

where $k \leq c_n(\eta)|\log \delta|^{n+1} + |\log \delta|$ *and* $c_n(\eta)$ *depends only on the spatial dimension* n *and* η.

Proof. Let $y \in D_2$ and $r > 0$ such that $B_r(y) \subset D_2$. For fixed $x \in D_1$ define

$$u(y) = \hat{G}(x,y) - \hat{G}_k(x,y) = \hat{G}(x,y) - \sum_{i=1}^k \hat{u}_i(x)\hat{v}_i(y).$$

Since $\hat{L}u = 0$, we obtain by the estimate

$$|u(y)|^2 \leq \frac{\tilde{c}_n}{\operatorname{vol} B_r(y)} \int_{B_r(y)} |u(z)|^2 \,\mathrm{d}z,$$

see [15, Thm. 8.17], together with (12) that

$$\begin{aligned} |\hat{G}(x,y) - \hat{G}_k(x,y)|^2 = |u(y)|^2 &\leq \frac{\tilde{c}_n}{\operatorname{vol} B_r(y)} \|\hat{G}(x,\cdot) - \hat{G}_k(x,\cdot)\|^2_{L^2(B_r(y))} \\ &\leq \frac{\tilde{c}_n}{\operatorname{vol} B_r(y)} \delta^2 \|\hat{G}(x,\cdot)\|^2_{L^2(B_{r(1+1/\eta)}(y)} \\ &= \frac{\tilde{c}_n (1+1/\eta)^n}{\operatorname{vol} B_{r(1+1/\eta)}(y)} \delta^2 \|\hat{G}(x,\cdot)\|^2_{L^2(B_{r(1+1/\eta)}(y))}. \end{aligned}$$

The assertion follows from the fact that the ratio

$$\frac{\|\hat{G}(x,\cdot)\|^2_{L^2(B_{r(1+1/\eta)}(y))}}{\operatorname{vol} B_{r(1+1/\eta)}(y)}$$

goes to $|\hat{G}(x,y)|^2$ as $r \to 0$, since $\hat{G}(x,\cdot)$ is continuous on D_2. □

Now that we known that $\hat{G}$ can be approximated on each admissible pair of domains (D_1, D_2), define $G_k(x,y)$ by

$$G_k(x,y) := \epsilon^{-1} e^{\phi_\epsilon(y)} \hat{G}_k(x,y) e^{-\phi_\epsilon(x)} = \sum_{i=1}^{k} \epsilon^{-1} e^{-\phi_\epsilon(x)} u_i(x) e^{\phi_\epsilon(y)} v_i(y).$$

Then G_k has the same degree of degeneracy as $\hat{G}_k$ and for the approximation error it holds that

$$\begin{aligned} |G(x,y) - G_k(x,y)| &= \epsilon^{-1} e^{\phi_\epsilon(y) - \phi_\epsilon(x)} |\hat{G}(x,y) - \hat{G}_k(x,y)| \\ &\leq \epsilon^{-1} e^{\phi_\epsilon(y) - \phi_\epsilon(x)} \delta |\hat{G}(x,y)| \\ &= \delta |G(x,y)|. \end{aligned}$$

Hence, we have derived the

Theorem 3.4. *Assume that* $\operatorname{dist}(D_1, D_2) \geq \eta \operatorname{diam} D_2$*. Then for any* $\delta > 0$ *there are functions* u_i*,* v_i*,* $i = 1, \ldots, k$*, such that* $G_k(x,y) := \sum_{i=1}^{k} u_i(x) v_i(y)$ *satisfies*

$$|G(x,y) - G_k(x,y)| \leq \delta |G(x,y)| \quad \textit{for all } x \in D_1 \textit{ and all } y \in D_2, \tag{14}$$

where $k \leq c_n(\eta) |\log \delta|^{n+1} + |\log \delta|$ *and* $c_n(\eta)$ *depends only on the spatial dimension* n *and* η*.*

4 $\mathcal{H}$-matrix approximation of SDFE matrices

Using a finite element discretization, $H^1_0(\Omega)$ is approximated by $V_h \subset H^1_0(\Omega)$, i.e., for all $v \in H^1_0(\Omega)$

$$\inf_{v_h \in V_h} \|v - v_h\|_{H^1} \to 0 \quad \text{for } h \to 0. \tag{15}$$

Usually, for V_h the set of piecewise linear functions defined on a polyhedrization $\mathcal{T}_h$ of Ω are used. We assume that $\varphi_i \geq 0$, $i \in I$, and $\sum_{i \in I} \varphi_i(x) = 1$ for all $x \in \Omega$. In agreement with the assumptions of Section 2, let $N = \dim V_h$ be the dimension and $\{\varphi_i\}_{i\in I}$ a basis of V_h, where $I := \{1, \ldots, N\}$ is used as an index set. The set $X_i := \operatorname{supp} X_i$ is the support of the ith basis function φ_i. It is well-known that the standard finite element discretization becomes unstable for convection-dominated problems. A stable alternative is the streamline diffusion finite-element method (SDFEM); see [26]. Its principle idea is to use $w_h := v_h + \alpha c \cdot \nabla v_h$ with an appropriately chosen parameter $\alpha > 0$ instead of $v_h \in V_h$ as trial functions. Since for C^0-elements $w_h \notin H^1$, the SDFEM is a non-conformal Petrov-Galerkin method. The variational formulation of (8) is then to find $u_h \in V_h$ such that

$$a(u_h, v_h + \alpha c \cdot \nabla v_h) = (f, v_h + \alpha c \cdot \nabla v_h) \quad \text{for all } v_h \in V_h,$$

where

$$a(u, v) = -\epsilon \sum_{\tau\in\mathcal{T}_h} (\Delta u, v)_\tau + (c \cdot \nabla u, v).$$

Here, $(\cdot,\cdot)_\tau$ denotes the scalar product in $L^2(\tau)$. Alternatively, one has the following variational formulation

$$a_h(u_h, v_h) = \ell_h(v_h) \quad \text{for all } v_h \in V_h, \tag{16}$$

where

$$a_h(u, v) = \epsilon(\nabla u, \nabla v) + (c \cdot \nabla u, \nabla v) + \sum_{\tau\in\mathcal{T}_h} \alpha_\tau(-\epsilon\Delta u + c \cdot \nabla u, c \cdot \nabla v)_\tau$$

and

$$\ell_h(v) = (f, v) + \sum_{\tau\in\mathcal{T}_h} \alpha_\tau(f, c \cdot \nabla v)_\tau.$$

By $J_1 : \mathbb{R}^N \to V_h$ and $J_2 : \mathbb{R}^N \to L^2(\Omega)$ we denote the natural bijections

$$J_1 x = \sum_{i\in I} x_i \varphi_i \quad \text{and} \quad J_2 x = \sum_{i\in I} x_i \psi_i,$$

where $\psi_i := \varphi_i + \alpha c \cdot \nabla\varphi_i$, $i \in I$. Using $\sum_{i\in I} \varphi_i = 1$, it follows with $e := (1, \ldots, 1)^T \in \mathbb{R}^N$ that

$$J_2 e = \sum_{i\in I} \psi_i = \sum_{i\in I} \varphi_i + \alpha c \cdot \nabla\varphi_i = 1 + \alpha c \cdot \sum_{i\in I} \nabla\varphi_i = 1 + \alpha c \cdot \nabla \sum_{i\in I} \varphi_i = 1.$$

Furthermore, from the non-negativity of φ_i we observe for $x \in \mathbb{R}^N$

$$|J_1 x| = |\sum_{i\in I} x_i \varphi_i| \le \sum_{i\in I} |x_i| \varphi_i = J_1|x|,$$

where $|x| \in \mathbb{R}^I$ denotes the vector with the components $|x_i|$, $i \in I$. Since $\nabla\varphi_i$ cannot be assumed to be non-negative, this estimate does not hold for J_2. However, we can derive a similar estimate

$$|J_2 x| \le J_2 \hat{x},$$

where $\hat{x} := |x| + c_g \|x\|_1 e$ and $c_g := 2\alpha \max_{i\in I} \|\nabla\varphi_i\|_2$. This estimate results from

$$\begin{aligned} |J_2 x| &\le \sum_{i\in I} |x_i| \left(\varphi_i + \alpha|c \cdot \nabla\varphi|\right) = J_2|x| + \alpha \sum_{i\in I} |x_i|(|c \cdot \nabla\varphi_i| - c \cdot \nabla\varphi_i) \\ &\le J_2|x| + 2\alpha \sum_{i\in I} |x_i||c \cdot \nabla\varphi_i| \le J_2|x| + c_g\|x\|_1 = J_2|x| + c_g\|x\|_1 J_2 e. \end{aligned}$$

In order to avoid technical complications, we consider a *quasi-uniform* and *shape-regular* triangulation. Hence, the step size $h := \max_{i\in I} \operatorname{diam} X_i$ fulfills

$$\operatorname{vol} X_i \geq c_U h^n, \quad i \in I \tag{17}$$

For quasi-uniform and shape-regular triangulations it is known ,see [21, Thm. 8.8.1], that there are constants $0 < c_{J,1} \leq c_{J,2}$ (independent of h and N) such that

$$c_{J,1}\|x\|_h \leq \|J_1 x\|_{L^2(\Omega)} \leq c_{J,2}\|x\|_h \qquad \text{for all } x \in \mathbb{R}^N, \tag{18}$$

where $\|\cdot\|_h$ is the naturally scaled Euclidean norm induced by the scalar product $\langle x, y\rangle_h = h^n \sum_{i\in I} x_i y_i$. Since J_1 is also a function from $\mathbb{R}^N$ to $H_0^1(\Omega)$, the adjoint $J_1^* \in L(H^{-1}(\Omega), \mathbb{R}^N)$ with respect to $\langle\cdot,\cdot\rangle_h$ is defined. We define the stiffness matrix $S \in \mathbb{R}^{N\times N}$ by $s_{ij} = a_h(\varphi_j, \varphi_i)$ and the Galerkin discretization of the inverse of L by

$$B = J_1^* L^{-1} J_2.$$

The matrix S is sparse, while B as well as S^{-1} are dense matrices.

Let $b = t \times s$ be an admissible block. The block B_b of the discrete inverse of L has the entries

$$B_{ij} = \int_\Omega \int_\Omega G(x,y)\varphi_i(x)\psi_j(y)\,\mathrm{d}y\,\mathrm{d}x, \quad (i,j) \in b.$$

Using the approximant G_k of G we define the approximant $B^{\mathcal{H}}$ for the entries in b by

$$B^{\mathcal{H}}_{ij} = \int_\Omega \int_\Omega G_k(x,y)\varphi_i(x)\psi_j(y)\,\mathrm{d}y\,\mathrm{d}x. \tag{19}$$

The rank of $B^{\mathcal{H}}_b$ is obviously bounded by k. It remains to estimate the approximation error. If $u \in \mathbb{R}^t$, $v \in \mathbb{R}^s$, from Lemma 3.2 we have

$$\begin{aligned}
|u^T(B_b - B^{\mathcal{H}}_b)v| &= |\int_\Omega \int_\Omega (G - G_k)(x,y)\; J_1 u(x)\; J_2 v(y)\,\mathrm{d}y\,\mathrm{d}x| \\
&\leq \int_\Omega \int_\Omega |G - G_k|(x,y)\; |J_1 u|(x)\; |J_2 v|(y)\,\mathrm{d}y\,\mathrm{d}x \\
&\leq \delta \int_\Omega \int_\Omega G(x,y)\; J_1|u|(x)\; J_2\hat{v}(y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \delta |u|^T B_b \hat{v} \leq \delta\|B_b\|_2\|u\|_2\,\|\hat{v}\|_2 \leq (c_g N + 1)\delta\|B_b\|_2\|u\|_2\|v\|_2,
\end{aligned}$$

because $\|\hat{v}\|_2 \leq \|v\|_2 + c_g\|v\|_1\|e\|_2 \leq \|v\|_2 + c_g\sqrt{N}\|v\|_2\sqrt{N}$. Hence,

$$\|B_b - B^{\mathcal{H}}_b\|_2 = \sup_{u,v} \frac{u^T(B_b - B^{\mathcal{H}}_b)v}{\|u\|_2\|v\|_2} \leq \delta(c_g N + 1)\|B_b\|_2. \tag{20}$$

The last estimate is an estimate of the relative approximation error on each admissible block b. For non-admissible blocks $b \in P$ we set $B^{\mathcal{H}}_b := B_b$ without approximation. Hence, the previous estimate holds for each block from the partition. In addition to blockwise estimates such estimates are required for the whole matrix. If we are interested in the Frobenius norm, estimates on each block A_b immediately lead to an estimate for A since

$$\|A\|_F^2 = \sum_{b\in P} \|A_b\|_F^2.$$

For the spectral norm this situation is a bit more difficult. We can however use the following lemma together with the structure of P.

Lemma 4.1. *Assume that the partition P is generated from $I \times I$ by recursively subdividing each block into a 2×2 block structure at most ℓ times. Furthermore, let $X, Y \in \mathbb{R}^{I \times I}$ such that $\|X_b\|_2 \leq \|Y_b\|_2$ for each block $b \in P$. Then it holds that*

$$\|X\|_2 \leq 2^\ell \|Y\|_2.$$

Proof. The assertion is proved by induction over the depth ℓ of the cluster tree. The estimate is trivially true if $\ell = 0$. Assume that the assertion holds for an $\ell \in \mathbb{N}$. Let

$$X = \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix} \quad \text{and} \quad Y = \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix}$$

have depth $\ell + 1$. Since X_{ij} and Y_{ij}, $i, j = 1, 2$, have depth ℓ, we know from the induction that

$$\|X_{ij}\|_2 \leq 2^\ell \|Y_{ij}\|_2, \quad i, j = 1, 2.$$

Observe that for matrices A_i, B_i satisfying $\|A_i\|_2 \leq \|B_i\|_2$, $i = 1, 2$, we have

$$\begin{aligned} \| \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} \|_2^2 &= \sup_{\|x\|_2=1} \|A_1 x\|_2^2 + \|A_2 x\|_2^2 \leq 2 \sup_{\|x\|_2=1} \max_{i=1,2} \|A_i x\|_2^2 \\ &= 2 \max_{i=1,2} \sup_{\|x\|_2=1} \|A_i x\|_2^2 = 2 \max_{i=1,2} \|A_i\|_2^2 \leq 2 \max_{i=1,2} \|B_i\|_2^2 \\ &= 2 \sup_{\|x\|_2=1} \max_{i=1,2} \|B_i x\|_2^2 \leq 2 \sup_{\|x\|_2=1} \|B_1 x\|_2^2 + \|B_2 x\|_2^2 = 2 \| \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} \|_2^2. \end{aligned}$$

Since $\|A\|_2 = \|A^T\|_2$ for any matrix A, we also obtain that $\| \begin{bmatrix} A_1 & A_2 \end{bmatrix} \|_2 \leq \sqrt{2} \| \begin{bmatrix} B_1 & B_2 \end{bmatrix} \|_2$. Hence,

$$\|X\|_2 = \| \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix} \|_2 \leq 2 \| \begin{bmatrix} 2^\ell Y_{11} & 2^\ell Y_{12} \\ 2^\ell Y_{21} & 2^\ell Y_{22} \end{bmatrix} \|_2 = 2^{\ell+1} \|Y\|_2,$$

which proves the assertion. □

From (20) and the last lemma we obtain the main result of this article.

Theorem 4.2. *Let $\delta > 0$ and let P be an admissible partition. Then there is $B^{\mathcal{H}} \in \mathcal{H}(P, k)$ with $k \leq c(\log N + |\log \delta|)^{n+1}$ such that*

$$\|B - B^{\mathcal{H}}\|_2 \leq \delta \|B\|_2$$

and c does neither depend on ϵ, N nor δ.

Proof. The assertion follows from (20) and the last lemma applied to P with $\ell \sim \log N$. □

The previous theorem shows that we are able to approximate the discrete inverse of L by $\mathcal{H}$-matrices. Our aim however is to prove that the inverse of the stiffness matrix S possesses this property. For this purpose we use the fact that S^{-1} can be approximated by $M_1^{-1} B M_2^{-1}$, where M_1 and M_2 denote the mass matrices for $\{\varphi_i\}$ and $\{\psi_i\}$, respectively. The last product, in turn, can be approximated by an $\mathcal{H}$-matrix. In [2] this technique has already been used to show that S^{-1} can be approximated up to the FE error. The same technique can be applied to convection-dominated problems without any changes. We remark that although this technique only provides an error estimate which is limited to the FE error, numerical experiments show that the inverse can be approximated with any prescribed precision.

We have already mentioned, see Theorem 2.3, that the existence of $\mathcal{H}$-matrix approximants to the factors of the LU decomposition follows from the existence of $\mathcal{H}$-matrix approximants to the inverse of a stiffness matrix. Hence, with Theorem 4.2 we obtain the existence result for the factors of the LU decomposition; see [5] for details.

5 Numerical experiments

In this section the practical influence of the parameter ϵ in (1) on the efficiency and accuracy of the $\mathcal{H}$-matrix approximation is investigated. We have seen that the $\mathcal{H}$-matrix inverse is important for theoretical purposes while the hierarchical LU decomposition is significantly faster in practice. Therefore, in the following tests we concentrate on the $\mathcal{H}$-LU decomposition.

For simplicity all tests are performed on a uniform triangulation of the unit square $\Omega := (0,1)^2$ in $\mathbb{R}^2$. Other (also three-dimensional) polyhedral domains do not cause any additional difficulties. We use piecewise linear elements for the discretization of $H_0^1(\Omega)$. Since for linear elements $\sum_\tau(\Delta u_h, c\cdot\nabla v_h)_\tau = 0$, equation (16) reads

$$\epsilon(\nabla u_h, \nabla v_h) + (c\cdot\nabla u_h, v_h) + \alpha(c\cdot\nabla u_h, c\cdot\nabla v_h) = (f, v_h) + \alpha(f, c\cdot\nabla v_h)$$

for all $v_h \in V_h$.

We report the results of applying the hierarchical LU decomposition with different precisions for both, irrotational convection and general convection. Another set of tests will employ low-precision LU decompositions for preconditioning. In each of the following cases the stiffness matrix S is built in the $\mathcal{H}$-matrix format. Then the usual decomposition algorithm[1], see [5], is applied to it with a relative rounding precision δ. Hence, the rank k is adaptively chosen and is therefore expected to vary among the blocks. All tests were carried out on an Athlon64 (2 GHz) PC with 4 GB of core memory.

5.1 Irrotational convection

In order to support our theory we first consider the irrotational vector field

$$c(x,y) = \begin{bmatrix} x - 1/2 \\ y - 1/2 \end{bmatrix}.$$

Table 1 shows the backward error

$$\mathcal{E} := \frac{\|A - L_\mathcal{H} U_\mathcal{H}\|_2}{\|A\|_2}$$

together with the CPU time for the computation of the factors $L_\mathcal{H}$ and $U_\mathcal{H}$ in the case $\delta = 0.1$. The columns labeled with MB show the memory requirements of $L_\mathcal{H}$ and $U_\mathcal{H}$. Table 2 contains the same quantities for the case $\delta = 1_{10}{-}4$.

Apparently, neither the efficiency nor the backward error $\mathcal{E}$ of the approximate LU decomposition depends on ϵ. As usual for hierarchical matrices, the complexity depends almost linearly on N. Changing the accuracy δ from $1_{10}{-}1$ to $1_{10}{-}4$ increases the computational effort by a factor of 1.7. The approximate LU decomposition appears to be backward stable.

[1] A software library can be obtained from `http://www.mathematik.uni-leipzig.de/~bebendorf/AHMED.html`.

	$\epsilon = 1_{10}-1$			$\epsilon = 1_{10}-7$			$\epsilon = 1_{10}-14$		
N	time	$\mathcal{E}$	MB	time	$\mathcal{E}$	MB	time	$\mathcal{E}$	MB
31 329	1.7s	$7.4_{10}-3$	21	1.8s	$5.5_{10}-3$	22	1.8s	$5.5_{10}-3$	22
63 001	4.1s	$1.0_{10}-2$	45	4.2s	$6.4_{10}-3$	48	4.2s	$6.4_{10}-3$	48
126 736	9.9s	$1.0_{10}-2$	95	10.3s	$5.8_{10}-3$	102	10.3s	$9.2_{10}-3$	102
254 016	21.8s	$6.9_{10}-3$	196	23.3s	$1.4_{10}-2$	218	23.2s	$9.7_{10}-3$	217
509 796	56.0s	$1.0_{10}-2$	414	60.2s	$2.7_{10}-2$	456	60.8s	$1.6_{10}-3$	465

Table 1: $\mathcal{H}$-*LU* with precision $\delta = 1_{10}-1$

	$\epsilon = 1_{10}-1$			$\epsilon = 1_{10}-7$			$\epsilon = 1_{10}-14$		
N	time	$\mathcal{E}$	MB	time	$\mathcal{E}$	MB	time	$\mathcal{E}$	MB
31 329	2.7s	$1.7_{10}-5$	31	2.7s	$2.3_{10}-5$	31	2.6s	$2.3_{10}-5$	31
63 001	6.6s	$2.0_{10}-5$	68	6.7s	$2.6_{10}-5$	71	6.7s	$2.5_{10}-5$	72
126 736	15.8s	$2.1_{10}-5$	148	16.5s	$1.4_{10}-5$	156	16.5s	$1.4_{10}-5$	156
254 016	37.6s	$2.1_{10}-5$	323	40.7s	$1.5_{10}-5$	349	40.3s	$1.5_{10}-5$	348
509 796	90.8s	$2.6_{10}-5$	699	106.4s	$2.0_{10}-5$	772	105.0s	$2.0_{10}-5$	770

Table 2: $\mathcal{H}$-*LU* with precision $\delta = 1_{10}-4$

5.2 Cyclic convection

In a second set of experiments we consider the cyclic vector field

$$c(x, y) = \begin{bmatrix} 0.5 - y \\ x - 0.5 \end{bmatrix}$$

as the convection term. Compared with the tests with irrotational convection, the numerical effort has only slightly increased. Tables 3 and 4 show that the complexity of the $\mathcal{H}$-*LU* decomposition is bounded for $\epsilon \to 0$.

	$\epsilon = 1_{10}-1$			$\epsilon = 1_{10}-7$			$\epsilon = 1_{10}-14$		
N	time	$\mathcal{E}$	MB	time	$\mathcal{E}$	MB	time	$\mathcal{E}$	MB
31 329	1.7s	$5.5_{10}-3$	21	1.8s	$1.4_{10}-2$	23	1.8s	$1.4_{10}-2$	23
63 001	4.1s	$4.7_{10}-3$	45	4.4s	$1.6_{10}-1$	51	4.4s	$1.8_{10}-1$	51
126 736	9.9s	$6.3_{10}-3$	95	10.5s	$1.2_{10}-2$	107	10.7s	$3.3_{10}-2$	108
254 016	22.6s	$4.9_{10}-3$	196	24.6s	$9.4_{10}-2$	224	24.6s	$2.6_{10}-2$	227
509 796	55.7s	$8.4_{10}-3$	415	59.7s	$5.2_{10}-2$	477	59.6s	$6.8_{10}-2$	474

Table 3: $\mathcal{H}$-*LU* with precision $\delta = 1_{10}-1$

According to [29], $-\operatorname{div} c \geq c_0$ with some $c_0 > 0$ is required for the coercivity of the bilinear form a_h from (16) on $V_h \times V_h$. Although for the cyclic field we have $\operatorname{div} c = 0$, the discretization seems to be stable. This is possibly due to numerical diffusion.

	$\epsilon = 1_{10}{-1}$			$\epsilon = 1_{10}{-7}$			$\epsilon = 1_{10}{-14}$		
N	time	$\mathcal{E}$	MB	time	$\mathcal{E}$	MB	time	$\mathcal{E}$	MB
31 329	2.7s	$2.8_{10}{-5}$	31	3.5s	$2.1_{10}{-5}$	37	3.5s	$2.1_{10}{-5}$	37
63 001	6.7s	$3.3_{10}{-5}$	70	9.1s	$1.9_{10}{-5}$	82	9.1s	$1.9_{10}{-5}$	83
126 736	15.9s	$2.9_{10}{-5}$	151	23.8s	$1.7_{10}{-5}$	183	24.1s	$1.7_{10}{-5}$	183
254 016	38.4s	$3.2_{10}{-5}$	328	62.9s	$1.3_{10}{-5}$	408	63.6s	$1.3_{10}{-5}$	409
509 796	93.8s	$2.9_{10}{-5}$	707	168.5s	$2.2_{10}{-5}$	908	168.7s	$2.2_{10}{-5}$	909

Table 4: $\mathcal{H}$-*LU* with precision $\delta = 1_{10}{-4}$

5.3 Employing the $\mathcal{H}$-*LU* decomposition as a preconditioner

For the preconditioning tests we consider the following vector field

$$c(x,y) = \begin{bmatrix} 0.5(1+\alpha) - y - \alpha x \\ x - \alpha y + 0.5(\alpha - 1) \end{bmatrix}$$

as the convection term. The parameter $\alpha > 0$ will be used to find out the importance of the presence of irrotational convection terms. The choice $\alpha = 0$ provides a purely cyclic vector field, while $\alpha > 0$ adds divergence. Table 5 shows the results obtained for $\alpha = 1_{10}{-1}$ and $\alpha = 1_{10}{-4}$. The rounding precision δ is chosen so that 50 iterations guarantee a relative residual error of at most $1_{10}{-4}$. The memory consumption of the preconditioner can be found in the columns 4 and 9, columns 3 and 8 contain the time for building the preconditioner, and in column 6 and 11 the time required by GMRES is reported.

	$\alpha = 1_{10}{-1}$					$\alpha = 1_{10}{-4}$				
N	δ	time	MB	#It	time	δ	time	MB	#It	time
31 329	$5.5_{10}{-1}$	1.1s	19	49	1.0s	$1.2_{10}{-1}$	1.8s	25	49	1.1s
63 001	$4.8_{10}{-1}$	2.7s	42	48	2.3s	$5.7_{10}{-2}$	4.4s	62	51	2.7s
126 736	$3.6_{10}{-1}$	8.3s	102	48	5.3s	$3.0_{10}{-2}$	13.6s	130	50	6.5s
254 016	$3.3_{10}{-1}$	17.7s	200	48	11.3s	$1.5_{10}{-2}$	34.9s	299	48	13.4s
509 796	$2.5_{10}{-1}$	57.6s	455	48	31.4s	$0.8_{10}{-2}$	135.3s	727	55	47.9s
1 020 100	$2.4_{10}{-1}$	115.3s	933	49	63.7s	$0.3_{10}{-2}$	363.8s	1641	46	97.6s

Table 5: $\mathcal{H}$-*LU* preconditioned GMRES with $\epsilon = 1_{10}{-14}$

From the numerical experiments above we conclude that the proposed preconditioner is robust with respect to the coefficient ϵ. Note that the proposed method does not rely on reordering the indices. The algorithms for the computation of the approximate *LU* decomposition were not adapted to the convection, the method rather adapts itself. Both, the storage requirements and the CPU time, scale almost linearly with the number of degrees of freedom N.

References

[1] M. Bebendorf: *Effiziente numerische Lösung von Randintegralgleichungen unter Verwendung von Niedrigrang-Matrizen.* dissertation.de, Verlag im Internet, 2001. ISBN 3-89825-183-7.

[2] M. Bebendorf and W. Hackbusch: *Existence of $\mathcal{H}$-Matrix Approximants to the Inverse FE-Matrix of Elliptic Operators with L^∞-Coefficients.* Numer. Math. 95, 1–28, 2003.

[3] M. Bebendorf: *Efficient Inversion of the Galerkin Matrix of General Second-Order Elliptic Operators with Nonsmooth Coefficients.* Math. Comp. 74, 1179–1199, 2005.

[4] M. Bebendorf: *Approximate Inverse Preconditioning of FE Systems for Elliptic Operators with non-smooth Coefficients.* Preprint 7/2004, Max-Planck-Institute MiS, Leipzig, accepted for publication in SIMAX.

[5] M. Bebendorf: *Why approximate LU decompositions of finite element discretizations of elliptic operators can be computed with almost linear complexity.* Preprint 8/2005, Max-Planck-Institute MiS, Leipzig.

[6] M. Benzi: *Preconditionig Techniques for Large Linear Systems: A Survey.* J. Comp. Phys. 182:418–477, 2002.

[7] M. Benzi and M. Tůma: *A sparse approximate inverse preconditioner for nonsymmetric linear systems*, SIAM J. Sci. Comput. 19, 968-994, 1998.

[8] M. Benzi and M. Tůma: *A comparative study of sparse approximative inverse preconditioners.* Appl. Numer. Math. 30: 305, 1999.

[9] J. Bey and G. Wittum: *Downwind numbering: Robust multigrid for convection diffusion problems.* Appl. Num. Math. 23, 177–192, 1997.

[10] G. Beylkin, R. Coifman and V. Rokhlin: *Fast wavelet transforms and numerical algorithms. I.* Comm. Pure Appl. Math. 44(2), 141–183, 1991.

[11] E. Chow: *A priori sparsity patterns for parallel sparse approximate inverse preconditioners*, SIAM J. Sci. Comput. 21, 1804–1822, 2000.

[12] E. Chow and Y. Saad: *Approximate inverse preconditioners via sparse-sparse iterations*, SIAM J. Sci. Comput. 19, 995–1023, 1998.

[13] W. Dahmen, S. Prössdorf and R. Schneider: *Wavelet approximation methods for pseudodifferential equations. II. Matrix compression and fast solution.* Adv. Comput. Math. 1(3-4), 259–335, 1993.

[14] W. Dahmen, S. Prössdorf and R. Schneider: *Wavelet approximation methods for pseudodifferential equations. I. Stability and convergence.* Math. Z. 215(4), 583–620, 1994.

[15] D. Gilbarg and N. S. Trudinger: *Elliptic partial differential equations of second order.* Springer-Verlag, Berlin, 2001. Reprint of the 1998 edition.

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

[17] L. Greengard and V. Rokhlin: *A new version of the fast multipole method for the Laplace equation in three dimensions.* Acta Numerica, 1997, pages 229–269. Cambridge Univ. Press, Cambridge, 1997.

[18] M. Grote and T. Huckle: *Parallel preconditioning with sparse approximate inverses.* SIAM J. Sci. Comput. 18, 838–853, 1997.

[19] M. Grüter and K.-O. Widman: *The Green function for uniformly elliptic equations.* Manuscripta Math. 37, 303–342, 1982.

[20] W. Hackbusch: *Multi-Grid Methods and Applications.* Springer, 1985.

[21] W. Hackbusch: *Theorie und Numerik elliptischer Differentialgleichungen.* B. G. Teubner, Stuttgart, 1996 - English translation: *Elliptic differential equations. Theory and numerical treatment.* Springer-Verlag, Berlin, 1992.

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

[23] W. Hackbusch and B. N. Khoromskij: *A sparse $\mathcal{H}$-matrix arithmetic. II. Application to multi-dimensional problems.* Computing 64, 21–47, 2000.

[24] W. Hackbusch and Z. P. Nowak: *On the fast matrix multiplication in the boundary element method by panel clustering.* Numer. Math. 54, 463–491, 1989.

[25] N. J. Higham: *Accuracy and stability of numerical algorithms.*, Second Edition, SIAM, Philadelphia, PA, 2002.

[26] T. J. R. Hughes and A. Brooks: *A multidimensional upwind scheme with no crosswind diffusion.* In T. J. R. Hughes, editor, Finite Element Methods for Convection Dominated Flows, AMD, vol. 34, ASME, pp. 19–35, New York, 1979.

[27] S. Le Borne: *$\mathcal{H}$-matrices for convection-diffusion problems with constant convection.* Computing 70, 261–274, 2003.

[28] A. Reusken: *Multigrid with matrix-independent transfer operators for a singular perturbation problem.* Computing 50, 199–211, 1993.

[29] H.-G. Roos, M. Stynes and L. Tobiska: *Numerical methods for singularly perturbed differential equations*, Springer Series in Computational Mathematics 24, Springer-Verlag, Berlin, 1996.

[30] J. W. Ruge and K. Stüben: *Algebraic multigrid.* in Multigrid Methods, edited by S. F. McCormick, p. 73, SIAM, 1987.

[31] Y. Saad: *Iterative Methods for Sparse Linear Systems.* PWS Publishing, Boston, 1996.

[32] M. Tabata: *A finite element approximation corresponding to the upwind differencing.* Memoirs of Numerical Mathematics 1:47–63, 1977.

[33] E. Tyrtyshnikov: *Mosaic-skeleton approximations.* Calcolo 33(1-2), 47–57 (1998), 1996.