Max-Planck-Institut
für Mathematik
in den Naturwissenschaften
Leipzig

Introduction to Hierarchical Matrices with Applications

by

Steffen Börm, Lars Grasedyck, and Wolfgang Hackbusch

Preprint no.: 18 2002

Introduction to Hierarchical Matrices with Applications

Steffen Börm[*], Lars Grasedyck[#], Wolfgang Hackbusch[*]
[*] Max-Planck-Institut *Mathematik in den Naturwissenschaften*
Inselstr. 22–26, D-04103 Leipzig, Germany
{sbo,wh}@mis.mpg.de
[#] Mathematisches Seminar, Universität zu Kiel, D-24098 Kiel, Germany
lgr@numerik.uni-kiel.de

Abstract

We give a short introduction to methods for the data-sparse approximation of matrices resulting from the discretisation of non-local operators occurring in boundary integral methods, as the inverses of partial differential operators or as solutions of control problems.

The result of the approximation will be so-called *hierarchical matrices* (or short $\mathcal{H}$ -matrices). These matrices form a subset of the set of all matrices and have a data-sparse representation. The essential operations for these matrices (matrix-vector and matrix-matrix multiplication, addition and inversion) can be performed in, up to logarithmic factors, optimal complexity.

We give a review of specialised variants of $\mathcal{H}$-matrices, especially of $\mathcal{H}^2$-matrices, and finally consider applications of the different methods to problems from integral equations, partial differential equations and control theory.

Keywords: Hierarchical matrices, data-sparse approximations, formatted matrix operations, fast solvers, Lyapunov equations, Riccati equations
Math Subject Classification: 65F05, 65F30, 65F50, 65N50

1 Introduction

1.1 Overview

$\mathcal{H}$-matrices are based on two observations:

- Integral operators can be efficiently treated by using separable expansions of the corresponding kernel functions (cf. [20, 26]).
- The inverse of an elliptic partial differential operator can be cast in the form of an integral operator by using Green's function as Schwartz kernel.

This article consists of seven sections: The current section gives a short overview of the basic ideas. The second section will establish *cluster trees* and *block partitions* that play a crucial role in the subsequent approximation schemes. The third section is concerned with operations on low-rank matrices, the basic building blocks of $\mathcal{H}$-matrices. In the following section, we introduce the original $\mathcal{H}$-matrices and a set of algorithms for performing basic algebraic operations on them. The fifth section is concerned with a specialisation of $\mathcal{H}$-matrices, namely uniform $\mathcal{H}$-matrices and $\mathcal{H}^2$-matrices, which can be used to significantly improve the performance and reduce memory requirements, especially for applications in the field of integral equations. Other specialised variants of $\mathcal{H}$-matrices are described in section six. The last section presents several applications of $\mathcal{H}$- and $\mathcal{H}^2$-matrices together with numerical results.

1.2 Model problem: Integral equation

Let us consider an integral operator of the form

$$\mathcal{L} : V \to V', \quad u \mapsto \left(x \mapsto \int_\Omega g(x,y)u(y)\,dy \right), \tag{1}$$

on a submanifold or subdomain Ω of $\mathbb{R}^d$ with a kernel function

$$g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}.$$

In typical applications, g is non-local, so, contrary to the treatment of differential operators, the finite element discretisation of the operator $\mathcal{L}$ does not lead to a sparse matrix. Due to the lack of sparsity, operations on the discrete matrix are prohibitively expensive.

There are different methods for avoiding the necessity of working with the full matrix: for some domains and some operators, it may be possible to diagonalise the matrix by applying the fast Fourier transform. In a more general setting, it may be possible to approximate the integral operator by means of a Wavelet basis.

$\mathcal{H}$-matrices are based on the fact that, at least for typical kernel functions $g(\cdot,\cdot)$, singularities only occur at the diagonal and the function is smooth everywhere else. In order to describe this property precisely, we introduce the notion of *asymptotic smoothness*: The kernel function $g(\cdot,\cdot)$ is called *asymptotically smooth*, if there are constants $C_{\mathrm{as1}}, C_{\mathrm{as2}} \in \mathbb{R}_{>0}$ satisfying

$$|\partial_x^\alpha \partial_y^\beta g(x,y)| \leq C_{\mathrm{as1}}(C_{\mathrm{as2}}\|x-y\|)^{-|\alpha|-|\beta|}|g(x,y)| \tag{2}$$

for all multi-indices $\alpha, \beta \in \mathbb{N}_0^d$ and all $x, y \in \mathbb{R}^d$ with $x \neq y$.

The idea of the *panel clustering method* (cf. [20]) is to exploit this smoothness in order to replace $g(\cdot,\cdot)$ by an approximation: if we fix $x \in \Omega$ and a subset Ω_y of Ω satisfying

$$\operatorname{dist}(x,\Omega_y) := \inf\{\|x-y\| \ : \ y \in \Omega_y\} \geq C \operatorname{diam}(\Omega_y),$$

we can replace $g(x,y)$ by its truncated Taylor expansion at a point $y_0 \in \Omega_y$ in order to get the *degenerate approximation*

$$\tilde{g}(x,y) = \sum_{\nu=1}^{k} p_\nu^x q_\nu(y). \tag{3}$$

If we use a collocation scheme, this translates into a *low-rank approximation* of part of the row corresponding to the collocation point x. By organising the sets Ω_y in a hierarchical way, a data-sparse approximation of the full matrix can be constructed.

A closer investigation shows that the coefficients p_ν^x can be used for a neighbourhood Ω_x of x without reducing the approximation properties. This leads to schemes based on low-rank approximations of blocks of the matrix (cf. [14, 26, 3]), that we will call $\mathcal{H}$-matrix schemes.

These methods can be refined by choosing different degenerate approximations, leading to multipole expansions, uniform $\mathcal{H}$-matrix or $\mathcal{H}^2$-matrix techniques.

1.3 Elliptic partial differential equations

Neglecting boundary conditions, we can represent the inverse of an elliptic partial differential operator by an integral operator of the form (1). This suggests that it will be possible to apply the $\mathcal{H}$-matrix technique to these inverses and even, since we assume that a finite element discretisation scheme will give us a good approximation of the infinite-dimensional operator, to store the inverse of the matrix resulting from a finite element discretisation in the form of an $\mathcal{H}$-matrix.

Since the "real" kernel function, i.e., Green's function, will depend on the (variable) coefficients and on the shape of the boundary in a complicated way, it is not feasible to discretise it directly in order to find an approximation of the inverse. Instead, we will introduce approximations of basic matrix operations like addition and multiplication in Subsection 4.4 and derive an approximative algorithm for the computation of the inverse of the FE stiffness matrix.

1.4 Matrix equations

The approximative matrix operations addition, multiplication and inversion introduced in Subsection 4.4 can replace the corresponding standard operations in many algorithms working on matrices:

We can approximate matrix functions like $\exp(L)$ by using the approximative matrix-matrix multiplication in combination with a Taylor expansion of the function or by using the approximative inversion in combination with Dunford-Cauchy integrals [6].

We can also use the approximative matrix arithmetic to improve the performance of Newton's method for solving nonlinear matrix equations like Riccati's equations from control theory.

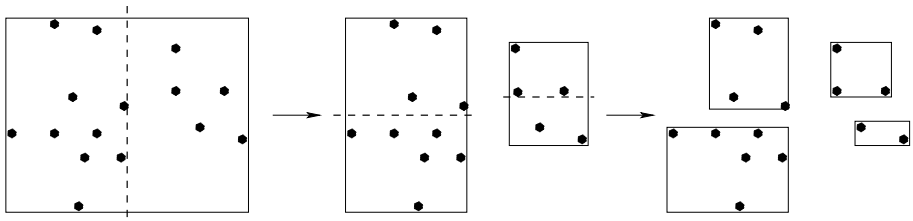


Figure 1: The bounding box to the left containing the points x_i is divided into two parts in x-direction. In the next step the new bounding boxes are divided in y-direction.

2 Construction of the cluster tree and block partition

While wavelet techniques can be employed to deal directly with problems in a continuum, $\mathcal{H}$-matrix techniques require a discrete subspace together with the finite element or boundary element basis $(\varphi_i)_{i\in I}$. The corresponding Ritz-Galerkin matrix (stiffness matrix) L is given by

$$L_{ij} = \langle \varphi_i, \mathcal{L}\varphi_j \rangle_{L^2}. \tag{4}$$

In the following, we identify subblocks $\tau \times \sigma \subset I \times I$ such that the submatrix $(L_{ij})_{i\in\tau, j\in\sigma}$ allows a low rank approximation. Since the number of possible subsets of $I \times I$ is considerably large we restrict ourselves to a small set of candidates τ, σ that are constructed and organised in a tree structure.

2.1 Cluster tree

Let $\mathcal{T}_I$ be a tree and denote by T_I the set of its nodes. $\mathcal{T}_I$ is called a *cluster tree* corresponding to an index set I, if the following conditions hold:

1. $T_I \subseteq \mathcal{P}(I) \setminus \{\emptyset\}$, i.e., each node of $\mathcal{T}_I$ is a subset of the index set I.
2. I is the root of $\mathcal{T}_I$.
3. If $\tau \in T_I$ is a leaf, then[1] $|\tau| \leq C_{\text{leaf}}$, i.e., the leaves consist of a relatively small number of indices.
4. If $\tau \in T_I$ is *not* a leaf, then it has two sons and is their disjoint union.

For each $\tau \in T_I$, we denote the set of its sons by $S(\tau) \subseteq T_I$. The leaves of the tree $\mathcal{T}_I$ are denoted by $\mathcal{L}_I$.

The restriction of $\mathcal{T}_I$ to binary trees serves only the purpose of simplifying the presentation of some steps of the algorithms. The extension to more general trees is straightforward.

The *support* of a cluster $\tau \in T_I$ is given by the union of the supports of the basis functions corresponding to its elements, i.e.,

$$\Omega_\tau := \bigcup_{i\in\tau} \Omega_i, \qquad \text{where } \Omega_i := \operatorname{supp} \varphi_i \text{ for all } i \in I.$$

Example 2.1 (Construction of cluster trees by bisection) *A simple method of building a cluster tree is based on geometry-based splittings of the index set. We associate each degree of freedom $i \in I$ with a suitable point $x_i \in \mathbb{R}^d$, e.g., the centre of the support of the corresponding basis function or the corresponding Lagrange point, if Lagrangian finite elements are used.*

Let $\{e_1, \ldots, e_d\}$ be an orthonormal basis of $\mathbb{R}^d$, e.g., the basis $\{e_x, e_y, e_z\}$ of the canonical unit vectors in 3D. The following algorithm will split a given cluster $\tau \subseteq I$ into two sons such that the points x_i are separated by a hyper-plane (see Figure 1).

```
procedure Split(τ, var τ1, var τ2);
begin
  { Choose a direction for geometrical splitting of the cluster τ }
  for j := 1 to d do begin
```

[1] $|\tau|$ denotes the number of elements in the set τ.

```
    α_j := min{⟨e_j, x_i⟩ : i ∈ τ};      { ⟨·,·⟩ is the ℝ^d Euclidean product }
    β_j := max{⟨e_j, x_i⟩ : i ∈ τ}
  end;
  j_max := argmax{β_j − α_j : j ∈ {1, ..., d}};
  { Split the cluster τ in the chosen direction }
  γ := (α_{j_max} + β_{j_max})/2;
  τ_1 := ∅; τ_2 := ∅;
  for i ∈ τ do
    if ⟨e_{j_max}, x_i⟩ ≤ γ then
      τ_1 := τ_1 ∪ {i}
    else
      τ_2 := τ_2 ∪ {i};
end
```

In theory each node of the tree is split into two sons until the cardinality of the node is 1*. In practice one should stop the splitting if the cardinality of a node is less or equal to a threshold parameter* C_{leaf}*. On typical machines, setting* $C_{\text{leaf}} = 32$ *leads to good performance.*

2.2 Admissibility condition

Next, we need an *admissibility condition* that allows us to check if a candidate $(\tau, \sigma) \in T_I \times T_I$ allows for a suitable low rank approximation.

If we assume asymptotically smooth kernels, this requirement will lead to an admissibility condition of the form

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

where $\eta \in \mathbb{R}_{>0}$ is some parameter controlling the trade-off between the number of admissible blocks, i.e., the algorithmic complexity, and the speed of convergence, i.e., the quality of the approximation (see, e.g., [20]).

In typical applications for unstructured grids, the computation of the diameter of a cluster and especially of the distance of two clusters will be too complicated or too time-consuming, so the "minimal" condition (5) will be replaced by a stronger variant, for example by using super-sets of Ω_τ and Ω_σ that are of a simpler structure.

Example 2.2 (Admissibility by Bounding Boxes) *A relatively general and practical admissibility condition for clusters in* $\mathbb{R}^d$ *can be defined by using* bounding boxes*: We define the canonical coordinate maps*

$$\pi_k : \mathbb{R}^d \to \mathbb{R}, \quad x \mapsto x_k,$$

for all $k \in \{1, \ldots, d\}$*. The bounding box for a cluster* $\tau \in T_I$ *is then given by*

$$Q_\tau := \prod_{k=1}^{d} [a_{\tau,k}, b_{\tau,k}], \quad \text{where} \quad a_{\tau,k} := \min(\pi_k \Omega_\tau) \quad \text{and} \quad b_{\tau,k} := \max(\pi_k \Omega_\tau).$$

Obviously, we have $\Omega_\tau \subseteq Q_\tau$*, so we can define the admissibility condition*

$$\min\{\operatorname{diam}(Q_\tau), \operatorname{diam}(Q_\sigma)\} \leq \eta \operatorname{dist}(Q_\tau, Q_\sigma) \tag{6}$$

that implies (5). We can compute the diameters and distances of the boxes by

$$\operatorname{diam}(Q_\tau) = \left(\sum_{k=1}^{d} (b_{\tau,k} - a_{\tau,k})^2 \right)^{1/2} \quad \text{and}$$

$$\operatorname{dist}(Q_\tau, Q_\sigma) = \left(\sum_{k=1}^{d} \left(\max(0, a_{\tau,k} - b_{\sigma,k})\right)^2 + \left(\max(0, a_{\sigma,k} - b_{\tau,k})\right)^2 \right)^{1/2}.$$

2.3 Block tree

The cluster tree can be used to define a *block tree* by forming pairs of clusters recursively. The block tree $\mathcal{T}_{I\times I}$ corresponding to a cluster tree $\mathcal{T}_I$ and an admissibility condition is constructed by the following procedure:

```
procedure BuildBlockTree(τ × σ);
begin
  if τ × σ is not admissible and |τ| > C_leaf and |σ| > C_leaf then begin
    S(τ × σ) := {τ' × σ' : τ' ∈ S(τ), σ' ∈ S(σ)};
    for τ' × σ' ∈ S(τ × σ) do BuildBlockTree(τ' × σ')
  end
  else
    S(τ × σ) := ∅
end
```

By calling this procedure with $\tau = \sigma = I$ and a parameter C_{leaf} that determines the minimal size of a cluster (see Definition 4.1), we create a block tree with root $I \times I$. The leaves of the block tree are denoted by $\mathcal{L}_{I\times I}$ and form a partition of $I \times I$.

The suitability of a block tree can be measured by

Definition 2.3 (Sparsity) *We define the sparsity constant C_{sp} of a block tree $\mathcal{T}_{I\times I}$ by*

$$C_{\text{sp}} := \max_{\tau\in T_I} |\{\sigma \in T_I \ ; \ \tau \times \sigma \in \mathcal{L}_{I\times I}\}|.$$

A block tree $\mathcal{T}_{I\times I}$ is called sparse *if $C_{\text{sp}} = \mathcal{O}(1)$ (does not depend on $|I|$).*

The constant C_{sp} is a measure for the sparsity of the block structure imposed by the partitioning $\mathcal{L}_{I\times I}$ of the product index set $I \times I$ (see Figure 2).

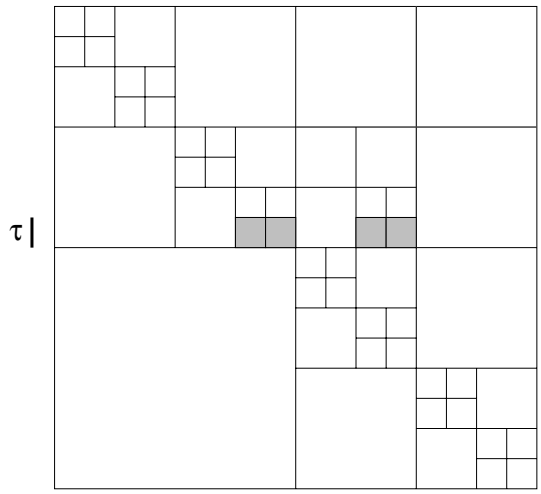


Figure 2: The maximum of 4 in the definition of C_{sp} is achieved by τ.

The complexity of algorithms for the creation of a cluster tree and block tree has been analysed in detail in [10]: for typical quasi-uniform grids, a "good" cluster tree can be created in $\mathcal{O}(n \log n)$ operations, the computation of the block tree can be accomplished in $\mathcal{O}(n)$ operations. Even for arbitrary grids where we only assume that the supports Ω_i of the corresponding basis functions are local and satisfy some weak assumptions, one can generate a sparse block tree in $\mathcal{O}(n \log n)$ operations (see [11, 13]). An alternative algorithm for constructing admissible block-partitions for integral equations is described in [3]. Further analysis concerning the approximation on graded meshes can be found in [16].

3 Rk -matrices

The basic building blocks for $\mathcal{H}$-matrices (defined in Section 4) are Rk -matrices which are a straight-forward representation of low rank matrices. These matrices form subblocks of the $\mathcal{H}$-matrix corresponding to subsets $\tau \times \sigma \subset I \times I$.

Definition 3.1 (Rk -matrix) *A matrix of the form*

$$R = AB^T, \qquad A \in \mathbb{R}^{\tau \times k}, B \in \mathbb{R}^{\sigma \times k}$$

is called an Rk *-matrix.*

Any matrix of rank at most k can be represented as an Rk -matrix and each Rk -matrix has at most rank k. Rk -matrices have some nice properties, e.g., only $k(n+m)$ numbers are needed to store an Rk -matrix.

3.1 Discretisation

In the $\mathcal{H}$-matrix representation of matrices, Rk -matrices will occur only as a representation of *admissible* blocks.

If $\mathcal{L}$ is a differential operator, we have $\operatorname{supp}(\mathcal{L}\varphi_j) \subseteq \operatorname{supp}\varphi_j$, so the matrix blocks corresponding to admissible pairs of clusters are zero.

The situation is more complicated if $\mathcal{L}$ is an integral operator of the type (1): let $\tau \times \sigma$ be an admissible pair of clusters. Without loss of generality, we may assume that $\operatorname{diam}(\Omega_\tau) \leq \operatorname{diam}(\Omega_\sigma)$.

In order to construct a rank-k approximation of the block $\tau \times \sigma$, we use an m-th order interpolation scheme[2] with interpolation points $(x_j^\tau)_{j=1}^k$ and corresponding Lagrange polynomials $(p_j^\tau)_{j=1}^k$ and approximate the original kernel function $g(\cdot,\cdot)$ by its interpolant

$$\tilde{g}(x,y) := \sum_{\iota=1}^{k} p_\iota^\tau(x) g(x_\iota^\tau, y). \tag{7}$$

Combining the asymptotical smoothness assumption (2) with standard interpolation error estimates, we get

$$|g(x,y) - \tilde{g}(x,y)| \leq C \left(C_{\mathrm{int}}\, C_{\mathrm{as2}}\, \frac{\operatorname{diam}(\Omega_\tau)}{\operatorname{dist}(\Omega_\tau, \Omega_\sigma)} \right)^m \|g\|_{\infty, \Omega_\tau \times \Omega_\sigma},$$

which combined with the admissibility condition (5) reads

$$|g(x,y) - \tilde{g}(x,y)| \leq C \left(C_{\mathrm{int}} C_{\mathrm{as2}}\, \eta \right)^m \|g\|_{\infty, \Omega_\tau \times \Omega_\sigma},$$

so if we choose $\eta < 1/(C_{\mathrm{int}} C_{\mathrm{as2}})$, we get *exponential* convergence of the interpolation if we increase the order m.

By replacing $g(\cdot,\cdot)$ by $\tilde{g}(\cdot,\cdot)$ in (4), we find

$$L_{ij} \approx \sum_{\iota=1}^{k} \int_\Omega p_\iota^\tau(x) \varphi_i(x)\, dx \int_\Omega g(x_\iota^\tau, y) \varphi_j(y)\, dy. \tag{8}$$

We define matrices $A \in \mathbb{R}^{\tau \times k}$ and $B \in \mathbb{R}^{\sigma \times k}$ by setting

$$A_{i\iota} := \int_\Omega p_\iota^\tau(x) \varphi_i(x)\, dx \quad \text{and} \quad B_{j\iota} := \int_\Omega g(x_\iota^\tau, y) \varphi_j(y)\, dy$$

and rewrite (8) as[3]

$$L|_{\tau \times \sigma} \approx AB^T,$$

so we have approximated $L|_{\tau \times \sigma}$ by an Rk -matrix.

[2] A scheme of this type can be easily constructed by extending one-dimensional interpolation techniques to the multi-dimensional case using tensor products. This leads to the relation $k = m^d$ between the rank of a block and the order of the interpolation.

[3] For a vector v and a subset $\tau \subset I$, $v|_\tau$ is the restriction to the vector $(v_j)_{j \in \tau}$, while for a matrix L and subsets $\tau, \sigma \subset I$ the notation $L|_{\tau \times \sigma}$ is used for the block $(L_{ij})_{i \in \tau, j \in \sigma}$.

Remark 3.2 (Collocation instead of Galerkin techniques) *This interpolation-based approach is not limited to Galerkin discretisations: In the case of a collocation scheme, the matrix entries* L_{ij}^{coll} *take the form*

$$L_{ij}^{\mathrm{coll}} := (\mathcal{L}\varphi_j)(c_i) = \int_\Omega g(c_i, y)\varphi_j(y)\,dy,$$

where $(c_i)_{i\in I}$ *is the family of collocation points. Replacing the kernel function* $g(\cdot,\cdot)$ *by its interpolant* $\tilde{g}(\cdot,\cdot)$*, we find*

$$L_{ij}^{\mathrm{coll}} \approx \sum_{\iota=1}^{k} p_\iota^\tau(c_i) \int_\Omega g(x_\iota^\tau, y)\varphi_j(y)\,dy$$

and can introduce

$$A'_{i\iota} := p_\iota^\tau(c_i)$$

in order to get

$$L^{\mathrm{coll}}|_{\tau\times\sigma} \approx A'B^T.$$

Remark 3.3 (Double layer potential) *The* double layer potential

$$\mathcal{L}^{\mathrm{DLP}} : V \to V', \quad u \mapsto \left(x \mapsto \int_\Gamma \langle \nabla_y g(x,y), n(y)\rangle u(y) dy\right)$$

for a 1*-codimensional submanifold* Γ *of* $\mathbb{R}^d$ *plays an important role in boundary element techniques. Since the effective kernel function depends on the normal vector of the manifold, it is not defined in* $\mathbb{R}^d$*, so we are not able to establish the estimate (2) directly.*

Instead, we replace $g(\cdot,\cdot)$ *by* $\tilde{g}(\cdot,\cdot)$ *from (7) and use the normal derivative of the result to get*

$$L_{ij}^{\mathrm{DLP}} \approx \sum_{\iota=1}^{k} \int_\Gamma p_\iota^\tau(x)\varphi_i(x)\,dx \int_\Gamma \langle \nabla_y g(x_\iota^\tau, y), n(y)\rangle \varphi_j(y)\,dy,$$

so we can again find a low-rank approximation

$$L^{\mathrm{DLP}}|_{\tau\times\sigma} \approx AB'^T$$

by setting

$$B'_{j\iota} := \int_\Gamma \langle \nabla_y g(x_\iota^\tau, y), n(y)\rangle \varphi_j(y)\,dy.$$

Remark 3.4 (Adaptive low-rank approximation) *The complexity can be significantly improved by using* adaptive *low-rank approximations instead of the analytically derived ones given above (cf. [1]).*

3.2 Matrix-vector multiplication

The matrix-vector multiplication $x \mapsto y := Rx$ of an $\mathbf{R}k$ -matrix $R = AB^T$ with a vector $x \in \mathbb{R}^\sigma$ can be done in two steps:

1. Calculate $z := B^T x \in \mathbb{R}^k$.
2. Calculate $y := Az \in \mathbb{R}^\tau$.

The transposed $R^T = BA^T$ can be treated analogously and the complexity of the matrix-vector multiplication is $\mathcal{O}(k(|\sigma| + |\tau|))$.

3.3 Truncation

A best approximation of an arbitrary matrix $M \in \mathbb{R}^{\tau\times\sigma}$ by an $\mathbf{R}k$ -matrix $\tilde{M} = \tilde{A}\tilde{B}^T$ (w.r.t. the spectral and Frobenius norm) can be computed using the (truncated) singular value decomposition as follows:

1. Calculate a singular value decomposition $M = U\Sigma V^T$ of M.
2. Set $\tilde{U} := [U_1 \dots U_k]$ (first k columns), $\tilde{\Sigma} := \operatorname{diag}(\Sigma_{11}, \dots, \Sigma_{kk})$ (first (largest) k singular values), $\tilde{V} := [V_1 \dots V_k]$ (first k columns).
3. Set $\tilde{A} := \tilde{U}\tilde{\Sigma} \in \mathbb{R}^{\tau\times k}$ and $\tilde{B} := \tilde{V} \in \mathbb{R}^{\sigma\times k}$.

We call $\tilde{M}$ a truncation of M to the set of $\mathbf{R}k$ -matrices. The complexity of the truncation is $\mathcal{O}((|\tau|+|\sigma|)^3)$.

If the matrix M is an $\mathbf{R}K$ -matrix $M = AB^T$ then the truncation can be computed in $\mathcal{O}(K^2(|\tau|+|\sigma|)+K^3)$ by the following procedure:

1. Calculate a truncated QR-decomposition $A = Q_A R_A$ of A, $Q_A \in \mathbb{R}^{\tau\times K}$, $R_A \in \mathbb{R}^{K\times K}$.
2. Calculate a truncated QR-decomposition $B = Q_B R_B$ of B, $Q_B \in \mathbb{R}^{\sigma\times K}$, $R_B \in \mathbb{R}^{K\times K}$.
3. Calculate a singular value decomposition $R_A R_B^T = U\Sigma V^T$ of $R_A R_B^T$.
4. Set $\tilde{U} := [U_1 \dots U_k]$ (first k columns), $\tilde{\Sigma} := \operatorname{diag}(\Sigma_{11}, \dots, \Sigma_{kk})$ (first (largest) k singular values), $\tilde{V} := [V_1 \dots V_k]$ (first k columns).
5. Set $\tilde{A} := Q_A\tilde{U}\tilde{\Sigma} \in \mathbb{R}^{\tau\times k}$ and $\tilde{B} := Q_B\tilde{V} \in \mathbb{R}^{\sigma\times k}$.

3.4 Addition

Let $R_1 = AB^T, R_2 = CD^T$ be $\mathbf{R}k$ -matrices. The sum

$$R_1 + R_2 = [A\ C][B\ D]^T$$

is an $\mathbf{R}2k$ -matrix. We define the *formatted addition* $\oplus$ of two $\mathbf{R}k$ -matrices as a best approximation (in the spectral and Frobenius norm) of the sum in the set of $\mathbf{R}k$ -matrices, which can be computed as in Section 3.3 with complexity $\mathcal{O}(k^2(|\tau| + |\sigma|) + k^3)$. The *formatted subtraction* $\ominus$ is defined analogously.

3.5 Multiplication

The multiplication of an $\mathbf{R}k$ -matrix $R = AB^T$ by an arbitrary matrix M from the left or right yields again an $\mathbf{R}k$ -matrix:

$$RM = AB^TM = A\,(M^TB)^T,$$
$$MR = MAB^T = (MA)\,B^T.$$

To calculate the product one has to perform the matrix-vector multiplication M^TB_i for the k columns $i = 1, \dots, k$ of B with the transposed of M or MA_i for the k columns $i = 1, \dots, k$ of A with the matrix M.

4 $\mathcal{H}$-matrices

Based on the cluster (binary) tree $\mathcal{T}_I$ and the block (quad-) tree $\mathcal{T}_{I\times I}$ we define the $\mathcal{H}$-matrix structure.

Definition 4.1 ($\mathcal{H}$-matrix) *Let $L \in \mathbb{R}^{I\times I}$ be a matrix and $\mathcal{T}_{I\times I}$ a block tree of $I \times I$ consisting of admissible and non-admissible leaves. Let $k \in \mathbb{N}$. L is called an* $\mathcal{H}$-matrix *of block-wise rank k, if for all admissible leaves $\tau \times \sigma \in \mathcal{T}_{I\times I}$*

$$\operatorname{rank}(L|_{\tau\times\sigma}) \le k,$$

i.e., each admissible subblock of the matrix can be represented as an $\mathbf{R}k$ -matrix while the non-admissible subblocks corresponding to leaves do not have to bear any specific structure.

Remark 4.2 *If $\tau\times\sigma$ is a non-admissible leaf of $\mathcal{T}_{I\times I}$, then either $|\tau| \le C_{\text{leaf}}$ or $|\sigma| \le C_{\text{leaf}}$ (see Subsection 2.3), which means that the rank is bounded by C_{leaf}.*

The storage requirements for an $\mathcal{H}$-matrix are $\mathcal{O}(nk\log(n))$ for the one- and two-dimensional block tree in [14] and [17]. The same bound holds for any $\mathcal{H}$-matrix based on a sparse block tree (see [10] and [11]).

4.1 Matrix-vector multiplication

Let $L \in \mathbb{R}^{I\times I}$ be an $\mathcal{H}$-matrix. To compute the matrix-vector product $y := y + Lx$ with $x, y \in \mathbb{R}^I$, we use the following procedure that performs the matrix-vector multiplication in each leaf of the block tree:

```
procedure MVM(L, τ × σ, x, var y);
begin
  if S(τ × σ) ≠ ∅ then
    for τ' × σ' ∈ S(τ × σ) do MVM(L, τ' × σ', x, y)
  else
    y|_τ := y|_τ + L|_{τ×σ} x|_σ; { unstructured or Rk-matrix }
end
```

The starting index sets are $\tau = \sigma = I$.

The complexity for the matrix-vector multiplication (sparse block tree) is $\mathcal{O}(nk\log(n))$ ([10], [11]). For some model problems the complexity can be estimated by exploiting the recursive structure as in [14, 17].

4.2 Addition

Let $L, L^{(1)}, L^{(2)} \in \mathbb{R}^{I\times I}$ be $\mathcal{H}$-matrices. The sum $L := L^{(1)} + L^{(2)}$ is an $\mathcal{H}$-matrix with block-wise rank $2k$. The *formatted sum* $\tilde{L} := L^{(1)} \oplus L^{(2)}$ is defined by using the formatted addition for the $\mathbf{R}k$ -subblocks and the standard addition for unstructured (full) matrices in the non-admissible leaves:

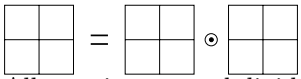
```
procedure Add(var L̃, τ × σ, L^(1), L^(2));
begin
  if S(τ × σ) ≠ ∅ then
    for τ' × σ' ∈ S(τ × σ) do Add(L̃, τ' × σ', L^(1), L^(2))
  else
    L̃|_{τ×σ} := L^(1)|_{τ×σ} ⊕ L^(2)|_{τ×σ} { unstructured or Rk-matrices }
end
```

Calling the procedure with $\tau = \sigma = I$, $\tilde{L} := 0$ yields $\tilde{L} = L^{(1)} \oplus L^{(2)}$.

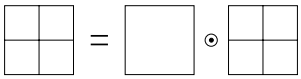
The complexity of the formatted addition (sparse block tree) is $\mathcal{O}(nk^2\log(n))$.

4.3 Multiplication

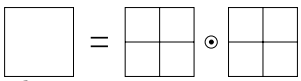
Let $L, L^{(1)}, L^{(2)} \in \mathbb{R}^{I\times I}$ be $\mathcal{H}$-matrices. The matrix $L := L + L^{(1)} \cdot L^{(2)}$ is (under moderate assumptions that are further investigated in [10] and [11]) an $\mathcal{H}$-matrix with block-wise rank $\mathcal{O}(k\log(n))$. The *formatted product* $\tilde{L} := L \oplus L^{(1)} \odot L^{(2)}$ is defined by using the formatted addition in the $\mathbf{R}k$ -subblocks. We distinguish between three cases:



All matrices are subdivided. The multiplication and addition is done in the subblocks.



The target matrix is subdivided and (at least) one of the factors is not subdivided. Then one of the factors has at most rank $\max\{k, C_{\text{leaf}}\}$. We have to compute the (low rank) product as in Subsection 3.5 and add the product to the target matrix.



The target matrix is not subdivided. This case will be treated in a separate procedure *MulAddRk*.

```
procedure MulAdd(var L̃, τ, ζ, σ, L^(1), L^(2));
begin
  if S(τ × ζ) ≠ ∅ and S(ζ × σ) ≠ ∅ and S(τ × σ) ≠ ∅ then
    { Case 1: All matrices are subdivided }
    for τ' ∈ S(τ), ζ' ∈ S(ζ), σ' ∈ S(σ) do
      MulAdd(L̃, τ', ζ', σ', L^(1), L^(2))
  else begin
    if S(τ × σ) ≠ ∅ then begin
      { Case 2: The target matrix is subdivided }
      Calculate the product L' := L^(1)|_{τ×ζ} L^(2)|_{ζ×σ}
      and add L' to L̃|_{τ×σ} { formatted addition in subblocks of τ × σ }
    end
    else begin
      { Case 3: The target matrix is not subdivided }
      MulAddRk(L̃, τ, ζ, σ, L^(1), L^(2))
    end
  end
end
```

Calling this procedure with $\tau = \zeta = \sigma = I, \tilde{L} := 0$ yields $\tilde{L} = L^{(1)} \odot L^{(2)}$.

To cover Case 3 we have to multiply two subdivided matrices, truncate the product to the set of $\mathbf{R}k$ -matrices and add the result to the target matrix. To do this we first calculate the products in the subblocks and truncate them to the set of $\mathbf{R}k$ -matrices. Afterwards all four $\mathbf{R}k$ -sub-matrices are added to the target matrix (extending them by zeros such that all matrices are of the same size) using the formatted addition.

```
procedure MulAddRk(var L̃, τ, ζ, σ, L^(1), L^(2));
begin
  if S(τ × ζ) = ∅ or S(ζ × σ) = ∅ then
    Calculate the product L' := L^(1)|_{τ×ζ} L^(2)|_{ζ×σ}
      and add L' to L̃|_{τ×σ} {formatted addition}
  else begin
    for each τ' ∈ S(τ), σ' ∈ S(σ) do begin
      Initialise L'_{τ',σ'} := 0;
      for each ζ' ∈ S(ζ) do MulAddRk(L'_{τ',σ'}, τ', ζ', σ', L^(1), L^(2));
      { L'_{τ',σ'} is smaller than L and extended by zeros }
      L̃ := L ⊕ Σ_{τ'∈S(τ)} Σ_{σ'∈S(σ)} L'_{τ',σ'}
    end
  end
end
```

The complexity for the formatted multiplication is $\mathcal{O}(nk^2 \log(n)^2)$ for the block trees from [14] and [17].

If $\mathcal{T}_{I\times I}$ is an arbitrary block tree that is sparse and *almost idempotent* (defined in [11]) then the complexity is again $\mathcal{O}(nk^2 \log(n)^2)$ (see [10] and [11]).

4.4 Inversion

The inverse of a 2×2 block-matrix can be computed by use of the Schur complement (see [14]) if the matrix is, e.g., positive definite. The exact sums and products are replaced by the formatted operations $\oplus, \odot$ and recursively one can define the *formatted inverse* $\tilde{L}$ of L by the following procedure:

```
procedure Invert(var L̃, τ, σ, L);
begin
  if S(τ × σ) = ∅ then
    Calculate the inverse L̃ := L^{-1} exactly        { Small matrix }
  else begin
    { S(τ) = {τ1, τ2}, S(σ) = {σ1, σ2}, L = [ L11  L12 ; L21  L22 ] }
```

```
    Invert(Y, τ1, σ1, L|τ1×σ1);
    S := L22 ⊖ (L21 ⊙ (Y ⊙ L12));
    Invert(L̃|τ2×σ2, τ2, σ2, S);
    L̃|τ1×σ1 := Y ⊕ (Y ⊙ (L12 ⊙ (L̃|τ2×σ2 ⊙ (L21 ⊙ Y))));
    L̃|τ1×σ2 := −Y ⊙ (L12 ⊙ L̃|τ2×σ2);
    L̃|τ2×σ1 := −L̃|τ2×σ2 ⊙ (L21 ⊙ Y)
  end
end
```

The starting index sets are $\tau = \sigma = I$. Note that two auxiliary matrices Y, S are needed in the procedure.

The complexity for the computation of the formatted inverse is the same as for the multiplication of two $\mathcal{H}$-matrices: $\mathcal{O}(nk^2 \log(n)^2)$. This is proven in [14] for a one-dimensional model problem but also holds for arbitrary block trees $\mathcal{T}_{I\times I}$ that are sparse and *almost idempotent* (see [10] and [11]).

5 $\mathcal{H}^2$-matrices

The matrix-vector multiplication for $\mathcal{H}$-matrices (cf. Subsection 4.1) has almost linear complexity. While this is sufficient for most applications, it is not optimal, so we will now consider improvements of the basic $\mathcal{H}$-matrix technique that lead to a better complexity. For some problems, even the optimal complexity of $\mathcal{O}(n)$ can be reached.

5.1 Uniform $\mathcal{H}$-matrices

Let us recall the approximation (8):

$$L_{ij} \approx \sum_{\iota=1}^{k} \int_\Omega p_\iota^\tau(x)\varphi_i(x)\,dx \int_\Omega g(x_\iota^\tau, y)\varphi_j(y)\,dy.$$

It was derived by interpolating the kernel function $g(\cdot,\cdot)$ in one argument, and the first argument x was chosen because of the smaller size of the support Ω_τ corresponding to the cluster τ. If Ω_σ had been smaller, we would have interpolated with respect to the second argument y.

We will now refine this approach along the lines described in [8] and introduce *uniform $\mathcal{H}$-matrices*, a subset of the general $\mathcal{H}$-matrices, that allow us to improve the complexity for the matrix-vector multiplication and of the storage requirements.

If the diameters of Ω_τ and Ω_σ are of comparable size, we can go one step further and interpolate in *both* variables, leading to the interpolant

$$\bar{g}(x,y) := \sum_{\iota=1}^{k_\tau}\sum_{\kappa=1}^{k_\sigma} g(x_\iota^\tau, x_\kappa^\sigma) p_\iota^\tau(x) p_\kappa^\sigma(y). \tag{9}$$

Since we are now interpolating in both arguments, we have to replace the standard admissibility condition (5) by the modified condition

$$\max\{\operatorname{diam}(\Omega_\tau), \operatorname{diam}(\Omega_\sigma)\} \le \eta \operatorname{dist}(\Omega_\tau, \Omega_\sigma) \tag{10}$$

in order to get approximation results similar to those derived for the $\mathcal{H}$-matrix technique. For most applications, the more restrictive new condition (10) and the original condition (5) will be equivalent (cf. [11]).

By replacing $g(\cdot,\cdot)$ by $\bar{g}(\cdot,\cdot)$ in (4), we get

$$L_{ij} \approx \sum_{\iota=1}^{k_\tau}\sum_{\kappa=1}^{k_\sigma} g(x_\iota^\tau, x_\kappa^\sigma) \int_\Omega p_\iota^\tau(x)\varphi_i(x)\,dx \int_\Omega p_\kappa^\sigma(y)\varphi_j(y)\,dy,$$

which can be rewritten in the form

$$L|_{\tau\times\sigma} \approx V_\tau S_{\tau,\sigma} V_\sigma^T \tag{11}$$

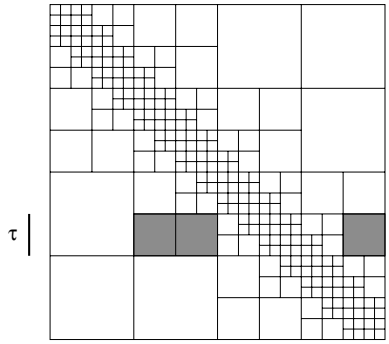


Figure 3: Blocks in the cluster row corresponding to τ

with

$$(V_\tau)_{i\iota} := \int_\Omega p_\iota^\tau(x)\varphi_i(x)\,dx, \quad (V_\sigma)_{j\kappa} := \int_\Omega p_\kappa^\sigma(y)\varphi_j(y)\,dy \quad \text{and} \tag{12}$$

$$(S_{\tau,\sigma})_{\iota\kappa} := g(x_\iota^\tau, x_\kappa^\sigma). \tag{13}$$

The important difference between the $\mathbf{R}k$ -representation used for standard $\mathcal{H}$-matrices and the representation (14) is the fact that the matrix V_τ depends only on the row cluster τ and the matrix V_σ depends only on the column cluster σ, while all the information about their interaction via the kernel function is stored in the, typically relatively small, matrix $S_{\tau,\sigma}$.

Definition 5.1 (Cluster basis) *Let $\mathbf{k} = (k_\tau)_{\tau\in T_I} \in \mathbb{N}^{T_I}$. A family $V = (V_\tau)_{\tau\in T_I}$ with $V_\tau \in \mathbb{R}^{\tau\times k_\tau}$ for each $\tau \in T_I$ is called a* cluster basis *with respect to the* rank distribution $\mathbf{k}$.

Definition 5.2 (Uniform $\mathcal{H}$-matrix) *Let $L \in \mathbb{R}^{I\times I}$ be a matrix and $\mathcal{T}_{I\times I}$ a block tree of $I \times I$ consisting of admissible and non-admissible leaves.*

Let V be a cluster basis with respect to a rank distribution $\mathbf{k}$.

L is called uniform $\mathcal{H}$-matrix *with respect to V and the coefficient family $(S_{\tau,\sigma})_{\tau\times\sigma\in\mathcal{L}_{I\times I}}$, if*

$$L|_{\tau\times\sigma} = V_\tau S_{\tau,\sigma} V_\sigma^T \tag{14}$$

holds for all $\tau \times \sigma \in \mathcal{L}_{I\times I}$.

Remark 5.3 (Subspace Property) *While the $\mathcal{H}$-matrices are not closed under addition (therefore we have to use the formatted addition mentioned above), the uniform $\mathcal{H}$-matrices corresponding to a fixed cluster basis V form a subspace of $\mathbb{R}^{I\times I}$.*

The decomposition (14) gives rise to a new algorithm for computing the matrix-vector product $y := Lx$. If we denote the set of all columns in a row corresponding to $\tau \in T_I$ (cf. Figure 3) by

$$R_\tau := \{\sigma \in T_I \ : \ \tau \times \sigma \in \mathcal{L}_{I\times I}\},$$

we find

$$y_i = (Lx)_i = \sum_{\tau\ni i}\left(\sum_{\sigma\in R_\tau} L|_{\tau\times\sigma}x|_\sigma\right)_i = \sum_{\tau\ni i}\left(\sum_{\sigma\in R_\tau} V_\tau S_{\tau,\sigma}V_\sigma^T x|_\sigma\right)_i = \sum_{\tau\ni i}\left(V_\tau \sum_{\sigma\in R_\tau} S_{\tau,\sigma}V_\sigma^T x|_\sigma\right)_i.$$

This representation of the result vector leads to the desired new algorithm: in a first step, we compute the vectors $\hat{x}_\sigma := V_\sigma^T x|_\sigma$, leading to the simplified equation

$$y_i = \left(\sum_{\tau\ni i} V_\tau \sum_{\sigma\in R_\tau} S_{\tau,\sigma}\hat{x}_\sigma\right)_i.$$

In the second step, we compute the inner sum

$$\hat{y}_\tau := \sum_{\sigma \in R_\tau} S_{\tau,\sigma} \hat{x}_\sigma.$$

In the last step, we compute the outer sum

$$y_i = \left(\sum_{\tau \ni i} V_\tau \hat{y}_\tau \right)_i.$$

The resulting algorithm takes the following form:

```
procedure UniformMVM(x, var y);
begin
  { Forward transformation }
  for σ ∈ T_I do x̂_σ := V_σ^T x|_σ;
  { Multiplication }
  for τ ∈ T_I do begin
    ŷ_τ := 0;
    for σ ∈ R_τ do ŷ_τ := ŷ_τ + S_{τ,σ} x̂_σ
  end;
  { Backward transformation }
  y := 0;
  for τ ∈ T_I do y|_τ := y|_τ + V_τ ŷ_τ
end
```

Remark 5.4 (Complexity) *In typical applications with "constant" rank $k_\tau = \min\{|\tau|, k_0\}$, the computation of the forward or backward transformation to a given vector can be computed in $\mathcal{O}(nk_0 \log n)$ operations. The multiplication requires only $\mathcal{O}(nk_0)$ operations.*

Although the complexity for the entire matrix-vector multiplication for uniform $\mathcal{H}$-matrices remains the same as in the case of $\mathcal{H}$-matrices, the new class of matrices has other advantages: they form a subspace, i.e., they can be added without truncation and the addition requires usually only $\mathcal{O}(nk_0)$ operations. If the cluster basis V, i.e., the subspace of the set of uniform $\mathcal{H}$-matrices, is fixed, only $\mathcal{O}(nk_0)$ units of memory are required to store an element of this class, therefore uniform $\mathcal{H}$-matrices can be used to significantly reduce memory requirements when working with more than one discretised operator, e.g., when storing resolvents $(A - z_i I)^{-1}$ for several shifts z_i.

5.2 $\mathcal{H}^2$-matrices

The asymptotic complexity of the matrix-vector multiplication for uniform $\mathcal{H}$-matrices is dominated by the forward and backward transformation. In order to come closer to the optimal complexity $\mathcal{O}(n)$, we have to find a way of improving these transformations. The method of doing this was introduced in [19] under the name of $\mathcal{H}^2$-matrices. While the former paper applies the Taylor expansion to approximate the kernel function, our presentation will be based on the interpolation approaches described in [8, 5].

Let us consider the case of an approximation of the type (9) with constant approximation order, i.e.,

$$\operatorname{span}\{p_\iota^{\tau_1} \ : \ \iota \in \{1, \ldots, k_{\tau_1}\}\} = \operatorname{span}\{p_\iota^{\tau_2} \ : \ \iota \in \{1, \ldots, k_{\tau_2}\}\}$$

holds for all $\tau_1, \tau_2 \in T_I$. This means that each polynomial $p_\iota^{\tau_1}$ corresponding to a cluster τ_1 can be expressed in the polynomial basis corresponding to any other cluster τ_2 and that we have constant rank $k_\tau = k_0$ (this implies $|\tau| \geq k_0$ for all clusters τ).

Since each $p_\iota^{\tau_2}$ is a Lagrange polynomial corresponding to an interpolation point $x_\iota^{\tau_2}$, we even find

$$p_\iota^{\tau_1}(x) = \sum_{\kappa=1}^{k_{\tau_2}} p_\iota^{\tau_1}(x_\kappa^{\tau_2}) p_\kappa^{\tau_2}(x).$$

Let us apply this equation to the case of a cluster τ with a son cluster τ'. For each $i \in \tau'$, we find

$$(V_\tau)_{i\iota} = \int_\Omega p_\iota^\tau(x) \varphi_i(x)\, dx = \sum_{\kappa=1}^{k_{\tau'}} p_\iota^\tau(x_\kappa^{\tau'}) \int_\Omega p_\kappa^{\tau'}(x) \varphi_i(x)\, dx = \sum_{\kappa=1}^{k_{\tau'}} p_\iota^\tau(x_\kappa^{\tau'}) (V_{\tau'})_{i\kappa},$$

i.e., we can express V_τ in terms of $V_{\tau'}$. We introduce the *transfer matrix* $B_{\tau',\tau}$ by setting

$$(B_{\tau',\tau})_{\kappa\iota} := p_\iota^\tau(x_\kappa^{\tau'})$$

and rewrite the above relation in the form

$$V_\tau|_{\tau\times k_\tau} = V_{\tau'} B_{\tau',\tau}. \tag{15}$$

This means that we only have to store V_τ for clusters without sons, while for all other clusters the usually much smaller transfer matrices $B_{\tau',\tau}$ are sufficient.

Definition 5.5 (Nested cluster basis) *A cluster basis V with respect to a rank distribution k is called* nested, *if there is a family $B = (B_{\tau',\tau})_{\tau\in T_I,\tau'\in S(\tau)}$ of* transfer matrices *satisfying*

$$V_\tau|_{\tau\times k_\tau} = V_{\tau'} B_{\tau',\tau}$$

for all $\tau \in T_I$ and $\tau' \in S(\tau)$.

Definition 5.6 ($\mathcal{H}^2$-matrix) *Let $L \in \mathbb{R}^{I\times I}$ be a uniform $\mathcal{H}$ -matrix with respect to a cluster basis V. The matrix L is called* $\mathcal{H}^2$-matrix, *if V is nested.*

The name "$\mathcal{H}^2$-matrix" is motivated by the fact that for this class of matrices, two hierarchies are involved: first the hierarchy of the clusters already exploited for $\mathcal{H}$-matrices, second the hierarchy of the cluster bases.

In addition to the fact that the memory requirements of $\mathcal{H}$-matrices are lower than those of $\mathcal{H}$- or even uniform $\mathcal{H}$-matrices, we can even speed up the forward and backward transformation: if τ is a cluster with sons $\{\tau_1,\ldots,\tau_s\}$, we have

$$V_\tau = \begin{pmatrix} V_{\tau_1} B_{\tau_1,\tau} \\ \vdots \\ V_{\tau_s} B_{\tau_s,\tau} \end{pmatrix} = \begin{pmatrix} V_{\tau_1} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & V_{\tau_s} \end{pmatrix} \begin{pmatrix} B_{\tau_1,\tau} \\ \vdots \\ B_{\tau_s,\tau} \end{pmatrix}, \tag{16}$$

and therefore

$$\hat{x}_\tau = V_\tau^T x|_\tau = \begin{pmatrix} B_{\tau_1,\tau}^T V_{\tau_1}^T & \ldots & B_{\tau_s,\tau}^T V_{\tau_s}^T \end{pmatrix} = \sum_{j=1}^s B_{\tau_j,\tau}^T V_{\tau_j}^T x|_{\tau_j} = \sum_{j=1}^s B_{\tau_j,\tau}^T \hat{x}_{\tau_j}.$$

This equation leads to the following recursive procedure that, given an input vector and a root cluster τ, will compute the coefficient vectors corresponding to all descendants of τ in the cluster tree:

```
procedure FastForward(τ, x, var x̂);
begin
  if S(τ) = ∅ then
    x̂_τ := V_τ^T x|_τ
  else begin
    x̂_τ := 0;
    for τ' ∈ S(τ) do begin
      FastForward(τ', x, x̂);
      x̂_τ := x̂_τ + B_{τ',τ}^T x̂_{τ'}
    end
  end
end
```

The backward transformation can, too, be rewritten as a recursive procedure:

```
procedure FastBackward(τ, var y, var ŷ);
begin
  if S(τ) = ∅ then
    y := V_τ ŷ_τ
  else
```

```
    for τ' ∈ S(τ) do begin
      ŷ_τ' := ŷ_τ' + B_τ',τ ŷ_τ;
      FastBackward(τ', y, ŷ)
    end
end
```

Remark 5.7 (Complexity) *The matrices $B_{\tau',\tau}$ are in $\mathbb{R}^{k_{\tau'} \times k_\tau}$, so the recursion steps in the new transformation algorithms require only $\mathcal{O}(k_{\tau'} k_\tau)$ operations, leading to a total of $\mathcal{O}(nk_0)$.*

In typical applications, the start of the iteration, i.e., the multiplication with V_τ in the leaves of $\mathcal{T}_I$, requires the same amount of work, leading to a total of $\mathcal{O}(nk_0)$.

The following algorithm results from combining the fast forward and backward transformations with the matrix-vector multiplication approach used for uniform matrices:

```
procedure FastMVM(x, var y);
begin
  { Recursive forward transformation }
  FastForward(I, x, x̂);
  { Multiplication }
  for τ ∈ T_I do begin
    ŷ_τ := 0;
    for σ ∈ R_τ do ŷ_τ := ŷ_τ + S_τ,σ x̂_σ
  end;
  { Recursive backward transformation }
  FastBackward(I, y, ŷ)
end
```

Remark 5.8 (Complexity) *By using the recursive algorithms, we have reduced the complexity of the forward and backward transformation to $\mathcal{O}(nk_0)$. In standard applications, the multiplication step requires the same amount of work, so the new matrix-vector multiplication algorithm has a complexity of $\mathcal{O}(nk_0)$.*

Remark 5.9 (Variable order) *Even the improved complexity of $\mathcal{O}(nk_0)$ is not optimal, since usually k_0 will be chosen to be proportional to $\log n$.*

In [25] a further refinement of $\mathcal{H}^2$-matrices is described: the order of approximation is no longer constant, but increases as the clusters become larger. This approach reduces the complexity for the matrix-vector multiplication and storage to the optimal class of $\mathcal{O}(n)$, while leaving the approximation properties intact, if the original kernel function $k(\cdot,\cdot)$ satisfies some additional assumptions.

5.3 Adaptive choice of the cluster basis

As soon as a cluster basis is fixed, the best uniform $\mathcal{H}$- or $\mathcal{H}^2$-approximation of a given matrix $L \in \mathbb{R}^{I \times I}$ in the Frobenius norm can be computed by solving the variational problem given by the projection equation

$$\langle V_\tau S_{\tau,\sigma} V_\sigma^T - L|_{\tau\times\sigma}, V_\tau X V_\sigma^T \rangle_F = 0$$

for all $X \in \mathbb{R}^{k_\tau \times k_\sigma}$. The solution is given by

$$S_{\tau,\sigma} = (V_\tau^T V_\tau)^{-1} V_\tau^T L|_{\tau\times\sigma} V_\sigma (V_\sigma^T V_\sigma)^{-1}. \tag{17}$$

If the bases V_τ and V_σ are orthogonal, this expression takes the simple form

$$S_{\tau,\sigma} = V_\tau^T L|_{\tau\times\sigma} V_\sigma. \tag{18}$$

This means that, in order to find a good approximation, we "only" have to determine a suitable cluster basis, preferably orthogonal, and compute the corresponding coefficient matrices by equation (17) or (18).

The *a priori* choice of a cluster basis will, in most applications, not lead to an optimal $\mathcal{H}^2$-representation of the operator. In some applications, e.g., when dealing with the inverses of finite element matrices or pseudo-differential operators, the theoretical construction of a suitable cluster basis will be much too complicated.

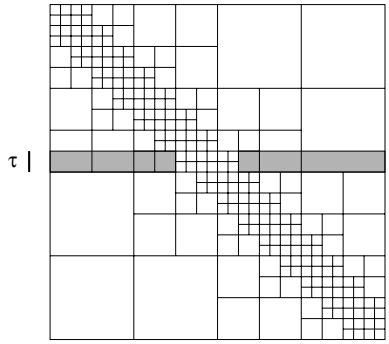


Figure 4: All blocks contributing to a cluster τ

In both cases, an algorithm for creating a cluster basis *a posteriori* is desirable. We will now describe an algorithm (introduced in [4]) for doing this.

In the case of uniform $\mathcal{H}$-matrices, we can use the following straightforward approach: in order to find an optimal rank-k-basis of a general matrix Y, we compute the singular value decomposition $U\Sigma V^T = Y$ of Y and use the first k columns of U as basis vectors. This is equivalent to choosing the eigenvectors corresponding to the k largest eigenvalues of the matrix $YY^T = U\Sigma^2 U^T$. Since a basis V_τ will in general be used by more than one block, we add up all blocks of the form $\tau \times \sigma$ with $\sigma \in R_\tau$ (recall Figure 3) to get

$$G_\tau := \sum_{\sigma \in R_\tau} L|_{\tau\times\sigma} L|_{\tau\times\sigma}^T.$$

The eigenvectors corresponding to the k largest eigenvalues of G_τ form the optimal basis V_τ.

The case of $\mathcal{H}^2$-matrices is more complicated: since the bases V_τ are required to be *nested*, they cannot be computed independently. Since the choice of a basis V_τ restricts the possible solutions for all ancestors of τ in the tree $\mathcal{T}_I$, blocks corresponding to these ancestors have to be considered in the computation of V_τ. To this end, we introduce the extended set

$$R_\tau^+ := \{\sigma \in T_I \ : \ (\exists \tau_0 \in T_I \ : \ \sigma \in R_{\tau_0} \wedge \tau \subseteq \tau_0)\}$$

representing all blocks corresponding to an ancestor of τ (cf. Figure 4).

We want to find an orthogonal basis V_τ that minimises the Frobenius error in all blocks

$$\sum_{\sigma \in R_\tau^+} \|L|_{\tau\times\sigma} - V_\tau V_\tau^T L|_{\tau\times\sigma}\|_F^2 = \sum_{\sigma \in R_\tau^+} \left(\|L|_{\tau\times\sigma}\|_F^2 - \|V_\tau V_\tau^T L|_{\tau\times\sigma}\|_F^2\right),$$

i.e., that maximises

$$\sum_{\sigma \in R_\tau^+} \|V_\tau^T L|_{\tau\times\sigma}\|_F^2 = \operatorname{tr}\left(V_\tau^T \left(\sum_{\sigma \in R_\tau^+} L|_{\tau\times\sigma} L|_{\tau\times\sigma}^T\right) V_\tau\right), \tag{19}$$

where $\operatorname{tr}(M) := \sum_{i\in I} M_{ii}$ denotes the trace of a matrix $M \in \mathbb{R}^{I\times I}$.

We can solve this problem by introducing the matrix

$$G_\tau := \sum_{\sigma \in R_\tau^+} L|_{\tau\times\sigma} L|_{\tau\times\sigma}^T,$$

computing its orthogonal diagonalisation (Schur decomposition)

$$Q_\tau D_\tau Q_\tau^T = G_\tau$$

and forming V_τ by picking the first k columns of Q_τ.

If τ has sons $\{\tau_1,\ldots,\tau_s\} = S(\tau) \neq \emptyset$, we have to ensure that the cluster bases are nested. We define

$$\bar{V}_\tau := \begin{pmatrix} B_{\tau_1,\tau} \\ \vdots \\ B_{\tau_s,\tau} \end{pmatrix} \tag{20}$$

and rewrite (16) in the form

$$V_\tau = \begin{pmatrix} V_{\tau_1} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & V_{\tau_s} \end{pmatrix} \bar{V}_\tau.$$

Combining this reformulation of the nestedness condition (15) with the maximisation problem (19), we find a new problem: now we have to find an orthogonal matrix $\bar{V}_\tau$ that maximises

$$\begin{aligned} &\operatorname{tr}\left(\bar{V}_\tau^T \left(\sum_{\sigma\in R_\tau^+} \begin{pmatrix} V_{\tau_1}^T & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & V_{\tau_s}^T \end{pmatrix} L|_{\tau\times\sigma} L|_{\tau\times\sigma}^T \begin{pmatrix} V_{\tau_1}^T & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & V_{\tau_s}^T \end{pmatrix}\right) \bar{V}_\tau\right) \\ = \operatorname{tr}\left(\bar{V}_\tau^T \left(\sum_{\sigma\in R_\tau^+} \begin{pmatrix} V_{\tau_1}^T L|_{\tau_1\times\sigma} \\ \vdots \\ V_{\tau_s}^T L|_{\tau_s\times\sigma} \end{pmatrix} \begin{pmatrix} V_{\tau_1}^T L|_{\tau_1\times\sigma} \\ \vdots \\ V_{\tau_s}^T L|_{\tau_s\times\sigma} \end{pmatrix}^T\right) \bar{V}_\tau\right) &= \operatorname{tr}\left(\bar{V}_\tau^T \left(\sum_{\sigma\in R_\tau^+} \bar{L}_{\tau,\sigma}\bar{L}_{\tau,\sigma}^T\right) \bar{V}_\tau\right) \end{aligned}$$

with

$$\widehat{L}_{\tau_i,\sigma} := V_{\tau_i}^T L|_{\tau_i\times\sigma} \quad \text{for } i\in\{1,\ldots,s\} \text{ and} \quad \bar{L}_{\tau,\sigma} := \begin{pmatrix} \widehat{L}_{\tau_1,\sigma} \\ \vdots \\ \widehat{L}_{\tau_s,\sigma} \end{pmatrix}. \tag{21}$$

This maximisation problem is again of the form discussed above and can therefore be solved by computing the orthogonal diagonalisation of

$$\bar{G}_\tau := \sum_{\sigma\in R_\tau^+} \bar{L}_{\tau,\sigma}\bar{L}_{\tau,\sigma}^T,$$

leading to the following algorithm:

```
procedure ComputeRowBasis(τ);
begin
  if S(τ) = ∅ then begin
    G_τ := 0;
    for σ ∈ R_τ^+ do G_τ := G_τ + L|_{τ×σ} L|_{τ×σ}^T;
    Compute the orthogonal diagonalisation Q_τ^T G_τ Q_τ = D;
    Form V_τ by copying the first k columns of Q_τ;
    for σ ∈ R_τ^+ do L̂_{τ,σ} := V_τ^T L|_{τ×σ}
  end
  else begin
    for τ' ∈ S(τ) doComputeRowBasis(τ');
    Ḡ_τ := 0;
    for σ ∈ R_τ^+ do Ḡ_τ := Ḡ_τ + L̄_{τ,σ} L̄_{τ,σ}^T;
    Compute the orthogonal diagonalisation Q_τ^T Ḡ_τ Q_τ = D;
    Form V̄_τ by copying the first k columns of Q_τ;
    Split V̄_τ into (B_{τ',τ})_{τ'∈S(τ)} according to (20);
    for σ ∈ R_τ^+ do L̂_{τ,σ} := V̄_τ^T L̄_{τ,σ}
  end
end
```

Remark 5.10 (Complexity) *For "constant" rank, i.e., for $k_\tau = \min\{|\tau|, k_0\}$, the algorithm requires $\mathcal{O}(n^2 k_0)$ operations. In the case of suitably varying rank, the algorithm reaches the optimal complexity $\mathcal{O}(n^2)$.*

Remark 5.11 (Application to $\mathcal{H}$-matrices) *The algorithm can, obviously, be applied to $\mathcal{H}$-matrices instead of general matrices. In this case, we can use the special structure in order to improve the complexity to $\mathcal{O}(nk_{\mathcal{H}}^2 \log(n))$, where $k_{\mathcal{H}}$ denotes the block-wise rank of the original $\mathcal{H}$-matrix.*

6 Alternative matrix formats

Besides $\mathcal{H}$-matrices and $\mathcal{H}^2$-matrices there are some specialised other matrix formats for certain applications, e.g.,

- wire-basket $\mathcal{H}$-matrices: if applications in the field of finite element matrices are considered, it is possible to combine $\mathcal{H}$-matrix ideas with local multi-grid solvers in order to reduce memory requirements [18];
- blended kernel approximation: for special geometries (e.g., surfaces of cylinders or bricks) the matrix of certain discretised integral operators takes the form of a tensor product of $\mathcal{H}$-matrices and circulant or Toeplitz matrices. This structure can be exploited by combining fast Fourier transforms and $\mathcal{H}$-matrix techniques to speed up the matrix vector multiplication [15];
- semi-explicit $\mathcal{H}$-matrices: the solution operator of a general elliptic problem can be expressed as a sum of the convolution operator corresponding to the fundamental solution and a pseudo-differential operator taking care of the boundary condition. The former operator can be treated by $\mathcal{H}$-matrix techniques while fast methods exist for the latter [21].

7 Applications

In this section, we will demonstrate how the $\mathcal{H}$-matrix techniques described above can be applied to problems from the fields of integral equations, differential equations and control theory.

7.1 Integral equations

In order to demonstrate the advantage of the $\mathcal{H}$-matrix approach, we consider the simple example of the discretisation of the single layer potential on the unit circle in two space dimensions using a Galerkin method with piecewise constant basis functions.

7.1.1 $\mathcal{H}$-matrix approach

The logarithmic kernel function will be approximated by the interpolation approach given in (7) using tensor product Chebyshev points and corresponding Lagrange polynomials.

We report the relative error[4] $\|L - \tilde{L}\|_2 / \|L\|_2$ in Table 1. The first column contains the number of degrees of freedom for each discretisation level, the following columns give the relative error. We observe that the error is bounded independently of the discretisation level and that it decreases very quickly when the interpolation order is increased.

n	1	2	3	4	5
1024	3.57-2	2.16-3	2.50-4	7.88-6	2.67-6
2048	3.58-2	2.19-3	2.51-4	7.86-6	2.69-6
4096	3.59-2	2.20-3	2.51-4	7.87-6	2.68-6
8192	3.59-2	2.20-3	2.52-4	7.76-6	2.67-6
16384	3.59-2	2.21-3	2.53-4	7.87-6	2.68-6

Table 1: Approximation error for the single layer potential

The time required for matrix vector multiplications is given in Table 2. It was measured on a SUN Enterprise 6000 machine using UltraSPARC II processors running at 248 MHz by taking the time required for

[4]The Euclidean norms are approximated by performing 100 steps of the power iteration.

100 matrix vector multiplications and dividing by 100. We can see that the complexity grows almost linearly with respect to the number of degrees of freedom and rather slowly with respect to the interpolation order.

n	1	2	3	4	5
1024	0.01	0.02	0.01	0.01	0.03
2048	0.02	0.04	0.03	0.05	0.07
4096	0.05	0.11	0.09	0.12	0.17
8192	0.12	0.24	0.19	0.26	0.39
16384	0.27	0.53	0.41	0.56	0.83
32768	0.57	1.15	0.90	1.23	1.90
65536	1.18	2.44	1.96	2.73	4.14
131072	2.45	5.18	4.30	5.89	8.98
262144	5.15	11.32	9.14	12.95	19.78
524288	10.68	23.81	19.62	28.02	43.57

Table 2: Time[sec] required for the matrix vector multiplication

Finally, let us consider the time required for building the $\mathcal{H}$-matrix representation of the discretised integral operator. It is given in Table 3 and was measured on the same machine. The integral of the Lagrange polynomials was computed by using an exact Gauss quadrature formula, while the integral of the kernel function was computed analytically. Once more we observe an almost linear growth of the complexity with respect to the number of degrees of freedom and a slow growth with respect to the interpolation order. Note that even on an old and quite slow processor like the 248 MHz UltraSPARC II, the boundary element matrix for more than half a million degrees of freedom can be approximated with an error less than 0.03% in less than half an hour.

n	1	2	3	4	5
1024	0.61	0.93	1.76	3.11	5.60
2048	1.25	2.03	3.85	7.04	12.94
4096	2.56	4.29	8.41	15.82	29.65
8192	5.25	9.16	18.10	35.31	66.27
16384	10.75	19.30	39.32	77.47	146.65
32768	22.15	40.83	85.16	169.16	324.36
65536	45.79	87.32	185.85	368.46	702.63
131072	92.64	180.73	387.63	788.06	1511.66
262144	189.15	378.20	854.75	1775.85	3413.45
524288	388.96	795.84	1743.66	3596.77	6950.55

Table 3: Time[sec] required for building the $\mathcal{H}$ -matrix

7.1.2 $\mathcal{H}^2$-matrix approach

In order to illustrate the advantages of $\mathcal{H}^2$-matrices when considering integral operators, we will now approximate the same kernel function as before by the symmetric interpolation approach given in (9), once more using tensor product Chebyshev points and the corresponding Lagrange polynomials.

n	1	2	3	4	5
1024	1.37-1	8.51-3	5.98-4	4.27-5	4.18-6
2048	1.37-1	8.56-3	5.98-4	4.29-5	4.19-6
4096	1.37-1	8.59-3	5.98-4	4.30-5	4.19-6
8192	1.37-1	8.60-3	5.98-4	4.31-5	4.19-6
16384	1.37-1	8.61-3	5.99-4	4.31-5	4.19-6

Table 4: $\mathcal{H}^2$-approximation error for the single layer potential

The relative approximation errors for the $\mathcal{H}^2$-matrix are reported in Table 4. As before in the $\mathcal{H}$-matrix case, the error is almost constant with respect to the number of degrees of freedom and decreases quickly when the order of the approximation is increased.

The results for the $\mathcal{H}^2$-matrix approximation are not as good as in the case of the $\mathcal{H}$-matrix for low ranks, but improve significantly as soon as the rank is increased.

n	1	2	3	4	5
1024	0.02	0.01	0.01	0.01	0.02
2048	0.04	0.03	0.03	0.03	0.07
4096	0.07	0.05	0.07	0.08	0.14
8192	0.15	0.10	0.17	0.16	0.29
16384	0.32	0.26	0.33	0.31	0.57
32768	0.66	0.49	0.66	0.60	1.12
65536	1.32	1.00	1.34	1.19	2.25
131072	2.68	2.00	2.75	2.50	4.77
262144	5.29	4.30	5.61	5.18	9.50
524288	10.72	8.26	10.91	9.99	18.57

Table 5: Time[sec] required for the $\mathcal{H}^2$ -matrix vector multiplication

In Table 5, we report the times required for the $\mathcal{H}^2$ matrix-vector multiplication. For the lowest rank, the $\mathcal{H}^2$-matrix technique requires more time than the $\mathcal{H}$-matrix approach, but even for order 2, the $\mathcal{H}^2$-matrix is significantly faster.

	1	2	3	4	5
1024	0.49	0.62	0.78	1.05	1.57
2048	0.98	1.22	1.49	2.11	3.32
4096	1.93	2.31	2.97	4.18	6.35
8192	3.94	4.64	6.16	8.47	12.99
16384	8.06	9.76	11.88	16.82	25.76
32768	15.86	18.75	24.25	33.53	51.97
65536	32.33	37.87	48.20	67.66	103.03
131072	66.71	75.46	96.63	139.71	208.18
262144	130.86	156.81	194.27	282.69	438.77
524288	264.06	307.09	390.56	548.59	839.58

Table 6: Time[sec] required for building the $\mathcal{H}^2$ -matrix

The Table 6 contains the times required for building the $\mathcal{H}^2$-matrix approximation of the integral operator. Obviously, the $\mathcal{H}^2$-matrix approach is much faster than the $\mathcal{H}$-matrix approach in *all* experiments. Especially in the case of higher order expansions, the performance of the $\mathcal{H}^2$-matrix technique is much better than that of the $\mathcal{H}$-matrix method.

Using $\mathcal{H}^2$-matrices, the boundary element matrix for more than half a million degrees of freedom can be approximated with an error less than 0.06% in less than seven minutes, even on a slow processor like the UltraSPARC II.

7.2 Elliptic partial differential equations

The model problem of the Poisson equation on the unit sphere is studied numerically in [10]. Some theoretical results will be presented in a forthcoming technical report [2]. For the convenience of the reader, we will give a short summary of the numerical tests from [10].

7.2.1 Simple model problem

We consider the Poisson equation

$$-\partial_x^2 u(x,y) - \partial_y^2 u(x,y) = f(x,y), \qquad (x,y) \in \Omega \tag{22}$$

on the unit square $\Omega = [0,1]^2$ with essential boundary condition $u = 0$ on the boundary $\Gamma := \partial\Omega$.

For the discretisation of the variational formulation of (22) we choose continuous piecewise linear nodal (Lagrangian) basis functions ϕ_i defined by a regular triangulation of $[0,1]^2$. The stiffness matrix

$$A_{i,j} := \int_\Omega \langle \nabla\phi_i, \nabla\phi_j \rangle, \qquad i,j \in I = \{1,\dots,n\} \tag{23}$$

and mass matrix

$$M_{i,j} := \int_\Omega \phi_i \phi_j, \qquad i,j \in I = \{1,\dots,n\} \tag{24}$$

are both sparse. Our aim is to compute an approximate inverse $\widetilde{A^{-1}}$ to A as in Subsection 4.4. The cluster tree $\mathcal{T}_I$ and the block tree $\mathcal{T}_{I\times I}$ are constructed as in Example 2.1 and Subsection 2.3 with admissibility condition (6) , $\eta = 0.8$.

The time needed to compute the approximate inverse $\widetilde{A^{-1}}$ to A is given in Table 7. If we fix the rank k then the required time grows almost linearly with respect to n ($9.3/1.9 = 4.9 \approx 4 = 512^2/256^2$).

	Number of degrees of freedom n				
k	32^2	64^2	128^2	256^2	512^2
1	9.3+0	6.8+1	4.3+2	1.9+3	9.3+3
2	9.7+0	8.0+1	5.0+2	2.7+3	1.4+4
3	1.1+1	9.7+1	6.4+2	3.7+3	2.0+4
4	1.2+1	1.2+2	8.3+2	5.1+3	2.6+4
5	1.3+1	1.4+2	1.1+3	6.6+3	3.5+4
10	1.9+1	2.6+2	2.4+3	1.5+4	
15	2.1+1	3.2+2	3.0+3	2.1+4	

Table 7: Time[sec] needed to compute the approximate inverse to A for increasing rank k.

Once an approximate inverse is computed, one can solve an equation of the form $Ax = b$ by $\tilde{x} := \widetilde{A^{-1}}b$. The time for one matrix-vector multiplication with the approximate inverse is reported in Table 8.

	Number of degrees of freedom n				
k	32^2	64^2	128^2	256^2	512^2
1	2.7-2	1.5-1	6.9-1	3.0+0	1.4+1
2	2.9-2	1.7-1	8.1-1	3.9+0	1.7+1
3	3.2-2	1.8-1	9.7-1	4.7+0	2.1+1
4	3.4-2	2.1-1	1.1-0	5.7+0	2.4+1
5	3.5-2	2.2-1	1.2-0	6.4+0	3.4+1
10	4.5-2	3.3-1	1.8-0	8.9+0	
15	5.0-2	3.9-1	2.3-0	1.2+1	

Table 8: Time[sec] needed for one matrix-vector multiplication for increasing rank k.

The relative error $\|A^{-1} - \widetilde{A^{-1}}\|/\|A^{-1}\|$ for the approximation of the inverse can be bounded by

$$\frac{\|A^{-1} - \widetilde{A^{-1}}\|}{\|A^{-1}\|} = \frac{\|(I - \widetilde{A^{-1}}A)A^{-1}\|}{\|A^{-1}\|} \le \|I - \widetilde{A^{-1}}A\|.$$

To measure $\|I - \widetilde{A^{-1}}A\|$ in the spectral norm, we perform 10 steps of the power iteration. The results are given in Table 9.

To compute a more accurate solution x of the equation $Ax = b$ one can use the linear iteration

$$x_0 := 0, \quad x_{i+1} := x_i - \widetilde{A^{-1}}(Ax_i - b),$$

that has a convergence rate of $\|I - \widetilde{A^{-1}}A\|$. Therefore, it suffices to compute an approximate inverse $\widetilde{A^{-1}}$ such that $\|I - \widetilde{A^{-1}}A\| < 1$. A solution of the linear equation $Ax = b$ can be gained up to any desired accuracy, but the discretisation error introduced by the choice of the subspace $\text{span}\{\phi_i : \ i \in I\}$ is a lower bound for the error of the solution to the continuous problem.

	Number of degrees of freedom n				
k	32^2	64^2	128^2	256^2	512^2
1	5.2-1	2.1-0	7.6-0	2.4+1	4.9+1
2	3.5-2	3.5-1	2.0-0	8.2-0	2.4+1
3	4.9-3	3.1-2	2.0-1	1.1-0	5.1-0
4	1.1-3	9.1-3	5.1-2	2.7-1	1.2-0
5	2.2-4	7.2-4	4.5-3	2.3-2	1.0-1
10	2.9-7	4.1-6	1.9-5	8.0-5	
15	8.4-13	5.4-10	2.2-9	2.8-8	

Table 9: Relative error $\|I - \widetilde{A^{-1}}A\|_2$ for the approximation of A^{-1} with increasing rank k.

7.2.2 Simple model problem with "jumping coefficients"

Changing the operator of equation (22) to

$$-\text{div}\,(\sigma(x,y)\nabla u(x,y)) = f(x,y), \qquad (x,y) \in \Omega \tag{25}$$

for some non-constant function σ does, as far as numerical tests indicate, not destroy the approximability of the inverse by an $\mathcal{H}$-matrix (cf., [10]). This is a rather surprising result: for the Poisson equation the singularity function is asymptotically smooth while here neither the singularity function nor its behaviour is known. For a theoretical investigation of the $\mathcal{H}$-matrix approximation of this problem, see [2].

7.3 Matrix functions

7.3.1 Matrix exponential function

Matrix functions like the matrix exponential can be computed effectively by use of the Dunford-Cauchy representation

$$\exp(A) = \frac{1}{2\pi i}\int_\Gamma \exp(z)(zI - A)^{-1}\text{dz} \tag{26}$$

of the operator. To do so, the matrices $zI - A$ have to be inverted. If the matrix A stems, e.g., from the discretisation of an elliptic partial differential equation, we can use the $\mathcal{H}$-matrix arithmetic to compute the inverse at certain points $z = z_j$ and compute the integral in (26) by some quadrature rule. This approach is investigated in [6], including numerical examples.

The Dunford-Cauchy representation can also be used for other operators [7].

An alternative way to compute the matrix exponential is using the truncated Taylor series with scaling. Let A be a matrix and $s \in \mathbb{N}$ such that $\|A\| < 2^{s-1}$. Then we find

$$\exp(A) = \exp(2^{-s}A)^{2^s} \approx \left(\sum_{i=0}^{k} \frac{1}{i!}(2^{-s}A)^i\right)^{2^s}.$$

Therefore one computes in a first step the matrix $B := \sum_{i=0}^{k} \frac{1}{i!}(2^{-s}A)^i$ by

```
procedure expsmall(var B, A, k);
begin
  B := k^{-1}A;
  for i := k - 1 downto 1 do B := i^{-1}(A ⊕ A ⊙ B);
  B := B + I
end
```

with calling parameters $B, (2^{-s}A)$. Afterwards the matrix B^{2^s} is gained by squaring s times,

$$B_0 := B, \quad B_{i+1} := B_i \odot B_i, \quad i = 1, \ldots, s-1 :$$

```
procedure exp(var B, A, k);
begin
  s := ⌈log2(‖A‖)⌉;
  A := 2^-s A;
  expsmall(B, A, k);
  for i := 1 to s do B := B ⊙ B
end
```

The complexity for the computation is $\mathcal{O}(\log(\|A\|)n\log(n)^2k^2)$ since $\log_2(\|A\|)$ formatted multiplications and additions of $\mathcal{H}$-matrices are involved.

7.3.2 Matrix sign function

Let S be a matrix with spectrum $\sigma(S)$ that does not intersect the imaginary axis. The matrix sign function $\text{sign}(S)$ is defined as the matrix function to $\text{sign} : \mathbb{C} \to \mathbb{C}, x + iy \mapsto \text{sign}(x)$. The sign of a number $x \neq 0$ is 1 if $x > 0$ and -1 if $x < 0$. This function plays an important role in the treatment of Riccati's equation.

A simple method to calculate the sign function of a matrix S is Newton's method applied to the equation $S^2 = I$, as it is described in [23].

Theorem 7.1 (Newton's method to calculate the matrix sign function) *Let $S \in \mathbb{R}^{n,n}$ be a matrix whose spectrum does not intersect the imaginary axis. Then the iteration*

$$S_0 := S, \qquad S_{i+1} := \frac{1}{2}(S_i + S_i^{-1}) \tag{27}$$

converges quadratically to the sign of S.

Replacing the exact addition and inversion by their formatted counterparts $\tilde{S}_{i+1} := \frac{1}{2}(\tilde{S}_i \oplus \widetilde{S_i^{-1}})$ yields an algorithm to compute the sign of an $\mathcal{H}$-matrix in the set of $\mathcal{H}$-matrices. To ensure a sufficient approximation of the matrix $\text{sign}(S)$ by $\tilde{S}_\infty := \lim_{i\to\infty} \tilde{S}_i$, one needs more assumptions concerning the structure of S. A numerical test involving the computation of $\tilde{S}_\infty$ will follow in the next subsection.

7.4 Lyapunov and Riccati equations

An algebraic matrix Riccati equation is an equation of the form

$$A^TX + XA - XFX + G = 0 \tag{28}$$

where the matrices A, F, G are given and X is the unknown solution.

Strategies for solving algebraic matrix Riccati equations (for matrices of a certain structure) are manifold. Basically, one can either try to solve the (nonlinear) equation (28) directly, or one can apply Newton's method to simplify the equation to a linear one. The latter results in a series of Lyapunov equations of the form

$$A^TX + XA + G = 0 \tag{29}$$

where the matrices A, G are given and X is the unknown solution.

7.4.1 Low rank case

If the matrix G in (29) is of (global) low rank and the spectra of A and $-A$ are disjoint, then the singular values of the solution X decay exponentially ([22], [9]). This means that the rank k needed to approximate the solution X up to a relative error of ε is $k = \mathcal{O}(-\log(\varepsilon))$.

If also the matrix F is of low rank, this extends to the case of the Riccati equation.

Consequently, an appropriate representation for X is that of an $\mathcal{R}k$ -matrix. Still one needs a method to compute X without losing the low rank structure.

Example 7.2 (Solution by use of the matrix exponential) *Let A be negative definite and G of low rank k_G. Then the solution X to (29) reads*

$$X = \int_0^\infty \exp(tA^T) G \exp(tA) dt.$$

The integral can be discretised and one has to compute the $\mathbb{R}^{k_G}$ -matrices $\exp(t_i A^T) G \exp(t_i A)$ for several $t_i \in (0, \infty)$.

This involves the computation of $\exp(t_i A)$ as in Subsection 7.3.1 and two products involving a low rank matrix that can be computed as in Subsection 3.5.

To solve the Riccati equation (28) we apply Newton's method and have to determine the iterates

$$X_{i+1} = \int_0^\infty \exp(t(A - FX_i)^T)(X_i F X_i + G) \exp(t(A - FX_i)) dt. \tag{30}$$

for a suitable initial guess X_0.

Example 7.3 (Solution by use of the matrix sign) *An algorithm to solve certain Riccati equations by use of the matrix sign function is presented in [23]: let $A \in \mathbb{R}^{n,n}$ be negative definite and $F, G \in \mathbb{R}^{n,n}$ symmetric positive semidefinite of low rank. Then the stabilising solution X of (28) satisfies*

$$\begin{bmatrix} N_{11} \\ N_{21} \end{bmatrix} X = - \begin{bmatrix} N_{12} \\ N_{22} \end{bmatrix}, \tag{31}$$

where the matrices $N_{11}, N_{12}, N_{21}, N_{22} \in \mathbb{R}^{n,n}$ are

$$\begin{bmatrix} N_{11} & N_{12} \\ N_{21} & N_{22} \end{bmatrix} := \operatorname{sign}\left(\begin{bmatrix} A^T & G \\ F & -A \end{bmatrix} \right) - \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}. \tag{32}$$

In the Lyapunov case $F = 0$ this simplifies to $X = N_{12}/2$.

In [12] it is proven that the matrix $\operatorname{sign}\left(\begin{bmatrix} A^T & G \\ F & -A \end{bmatrix} \right)$ consists of low rank structures plus the identity. The computation of $\operatorname{sign}(\dots)$ is done as in Subsection 7.3.2. The solution of (31) simplifies essentially if N_{11} is regular: $X = -N_{11}^{-1} N_{12}$. The exact inversion and multiplication are replaced by the formatted $\mathcal{H}$-matrix arithmetics.

To test the approximability of all the matrices appearing in (27) we solve the autonomous linear quadratic optimal control problem for the heat equation in one dimension. This has already been investigated in [24] but there the low rank structure of the solution was not exploited.

Example 7.4 (Control of the heat equation) *We consider the linear quadratic optimal control problem of the one-dimensional heat flow: The goal is to minimise*

$$J(u) := \int_0^\infty \left(y(t)^2 + u(t)^2 \right) dt \tag{33}$$

for $u \in L_2(0, \infty)$ where y is defined via the differential equation

$$\begin{array}{lcll} \frac{\partial}{\partial t} x(t,\xi) & = & \frac{\partial^2}{\partial \xi^2} x(t,\xi) + b(\xi) u(t), & \xi \in (0,1), t \in (0,\infty), \\ x(t,\xi) & = & 0, & \xi \in \{0,1\}, t \in (0,\infty), \\ x(0,\xi) & = & x_0(\xi), & \xi \in (0,1), \\ y(t) & = & \int_0^1 c(\xi) x(t,\xi) d\xi, & t \in (0,\infty). \end{array}$$

The starting value $x_0 \in L_2(0,1)$ is given (not important here) and b, c are

$$b(\xi) := \begin{cases} 1 & \xi \in (0.2, 0.3) \\ 0 & \textit{otherwise} \end{cases}, \qquad c(\xi) := \begin{cases} 1 & \xi \in (0.2, 0.3) \\ 0 & \textit{otherwise} \end{cases}.$$

The differential equation is discretised by finite differences on a uniform mesh of $(0,1)$ with n inner grid-points and mesh width h. If we define the matrices

$$A_{ij} := \begin{cases} 2h^{-2} & i = j \\ -h^{-2} & |i - j| = 1 \\ 0 & \textit{otherwise} \end{cases}, \qquad B_{i1} := \begin{cases} 1 & ih \in [0.2, 0.3] \\ 0 & \textit{otherwise} \end{cases},$$

$$C_{1j} := \int_{0.2}^{0.3} \phi_j(x)dx, \qquad i,j \in \{1,\dots,n\},$$

where ϕ_i denotes the i-th Lagrange basis function for the interpolation, then the minimising discrete control u is

$$u(t) = -B^T X x(t), \quad t \in (0,\infty),$$

where $X \in \mathbb{R}^{n,n}$ is the unique, nonnegative symmetric solution of the algebraic matrix Riccati equation

$$A^T X + XA - XFX + G = 0$$

for the matrices $F := BB^T$ and $G := C^T C$.

The matrices F, G are of rank one. Using the algorithm described in Example 7.3, we can compute an approximate solution $\tilde{X}$ to (28). The results can be seen in Table 10, where we present the relative error $\varepsilon := \|\tilde{X} - X\|_2 / \|X\|_2$ for increasing rank k and n degrees of freedom.

rank k	number of degrees of freedom n				
	256	1024	4096	16384	65536
k=1	1.5-1	1.3-1	2.5-0	divergent	divergent
k=2	2.6-4	4.2-4	1.2-3	5.6-4	6.7-4
k=3	1.2-5	1.3-5	1.5-5	2.3-5	3.9-5
k=4	9.1-8	1.1-7	1.0-6	1.8-6	6.2-7
k=5	4.6-9	1.1-8	1.5-8	3.0-8	3.1-8
k=6	3.7-10	2.4-10	4.9-10	5.9-10	1.7-9
iteration	14	17	20	23	26
time($k=2$)	8.5	67	462	3033	18263

Table 10: The last but one row shows the number of Newton steps to compute sign and the last row contains the time[sec] needed to compute the rank $k = 2$ solution on a Sun Quasar with 450 MHz.

7.4.2 $\mathcal{H}$-matrix case

If the matrix G in (29) is an $\mathcal{H}$-matrix with block-wise rank k_G, one can prove under moderate assumptions (see [12] for details and a proof) that the solution X can be approximated in the $\mathcal{H}$-matrix format with slightly increased block-wise rank k_X.

If also the matrix F is of low rank then this extends to the case of the Riccati equation.

Consequently, an appropriate representation for X is that of an $\mathcal{H}$-matrix. An approximate solution $\tilde{X}$ can be computed as in Example 7.2 or Example 7.3.

References

[1] M. Bebendorf, *Approximation of boundary element matrices*, Numer. Math., 86 (2000), pp. 565–589.

[2] M. Bebendorf and W. Hackbusch, *Existence of $\mathcal{H}$-matrix approximants to the inverse FE-matrix of elliptic operators with L^∞-coefficients*, Tech. Rep. to appear, Max Planck Institute for Mathematics in the Sciences, 2002.

[3] M. Bebendorf and S. Rjasanov, *Adaptive low-rank approximation of collocation matrices*, Tech. Rep. 39, Universität Saarbrücken, 2001.

[4] S. Börm and W. Hackbusch, *Data-sparse approximation by adaptive $\mathcal{H}^2$-matrices*, Tech. Rep. 86, Max Planck Institute for Mathematics in the Sciences, 2001.

[5] ——, *$\mathcal{H}^2$-matrix approximation of integral operators by interpolation*, Tech. Rep. 104, Max Planck Institute for Mathematics in the Sciences, 2001.

[6] I. Gavrilyuk, W. Hackbusch, and B. Khoromskij, *$\mathcal{H}$-matrix approximation for the operator exponential with applications*, Tech. Rep. 42, Max Planck Institute for Mathematics in the Sciences, 2000.

[7] ——, *$\mathcal{H}$-matrix approximation for elliptic solution operators on cylindric domains*, East-West J. Numer. Math., 9 (2001), pp. 25–58.

[8] K. Giebermann, *Multilevel approximation of boundary integral operators*, Computing, 67 (2001), pp. 183–207.

[9] L. Grasedyck, *Singular value bounds for the Cauchy matrix and solutions to Sylvester equations*, Tech. Rep. 13, University Kiel, 2001.

[10] ——, *Theorie und Anwendungen Hierarchischer Matrizen*, PhD thesis, Universität Kiel, 2001.

[11] L. Grasedyck and W. Hackbusch, *Construction and arithmetics of $\mathcal{H}$-matrices*, Tech. Rep. to appear, Max Planck Institute for Mathematics in the Sciences, 2002.

[12] L. Grasedyck, W. Hackbusch, and B. Khoromskij, *Application of $\mathcal{H}$-matrices in control theory*, Tech. Rep. to appear, Max Planck Institute for Mathematics in the Sciences, 2002.

[13] L. Grasedyck, W. Hackbusch, and S. LeBorne, *Adaptive refinement and clustering of $\mathcal{H}$-matrices*, Tech. Rep. 106, Max Planck Institute of Mathematics in the Sciences, 2001.

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

[15] W. Hackbusch and B. Khoromskij, *Blended kernel approximation in the $\mathcal{H}$-matrix techniques*, Tech. Rep. 66, Max Planck Institute for Mathematics in the Sciences, 2000.

[16] ——, *$\mathcal{H}$-matrix approximation on graded meshes*, in The Mathematics of Finite Elements and Applications X, J. R. Whiteman, ed., Elsevier, 2000, pp. 307–316.

[17] ——, *A sparse matrix arithmetic based on $\mathcal{H}$-matrices. Part II: Application to multi-dimensional problems*, Computing, 64 (2000), pp. 21–47.

[18] ——, *Towards $\mathcal{H}$-matrix approximation of linear complexity*, in Problems and Methods in Mathematical Physics. The Siegfried Prössdorf Memorial Volume, I. G. J. Elschner and B. Silbermann, eds., Birkhäuser, 2001, pp. 194–220.

[19] W. Hackbusch, B. Khoromskij, and S. Sauter, *On $\mathcal{H}^2$-matrices*, in Lectures on Applied Mathematics, H. Bungartz, R. Hoppe, and C. Zenger, eds., 2000, pp. 9–29.

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

[21] B. Khoromskij, *Data-sparse approximate inverse in elliptic problems: Green's function approach*, Tech. Rep. 79, Max Planck Institute for Mathematics in the Sciences, 2001.

[22] T. Penzl, *Eigenvalue decay bounds for solutions of Lyapunov equations: the symmetric case*, Systems and Control Letters, 40 (2000), pp. 139–144.

[23] J. D. Roberts, *Linear model reduction and solution of the algebraic Riccati equation by use of the sign function*, Internat. J. Control, 32 (1980), pp. 677–687.

[24] J. I. G. Rosen and C. Wang, *A multilevel technique for the approximate solution of operator Lyapunov and algebraic Riccati equations*, Siam J. Numer. Anal, 32 (1995), pp. 514–541.

[25] S. Sauter, *Variable order panel clustering*, Computing, 64 (2000), pp. 223–261.

[26] E. Tyrtyshnikov, *Mosaic-skeleton approximation*, Calcolo, 33 (1996), pp. 47–57.