Max-Planck-Institut
für Mathematik
in den Naturwissenschaften
Leipzig

Low-Rank Kronecker Product Approximation to Multi-Dimensional Nonlocal Operators. Part II. HKT Representation of Certain Operators

(revised version: September 2005)

by

Wolfgang Hackbusch and Boris N. Khoromskij

Preprint no.: 30 2005

Low-Rank Kronecker Product Approximation to Multi-Dimensional Nonlocal Operators. Part II. HKT Representation of Certain Operators

Wolfgang Hackbusch and Boris N. Khoromskij
Max-Planck-Institute for Mathematics in the Sciences,
Inselstr. 22-26, D-04103 Leipzig, Germany.
{wh, bokh}@mis.mpg.de

Abstract

This article is the second part continuing Part I [16]. We apply the $\mathcal{H}$-matrix techniques combined with the Kronecker tensor-product approximation to represent integral operators as well as certain functions $F(A)$ of a discrete elliptic operator A in a hypercube $(0,1)^d \in \mathbb{R}^d$ in the case of a high spatial dimension d. We focus on the approximation of the operator-valued functions $A^{-\sigma}$, $\sigma > 0$, and $\mathrm{sign}(A)$ for a class of finite difference discretisations $A \in \mathbb{R}^{N\times N}$. The asymptotic complexity of our data-sparse representations can be estimated by $\mathcal{O}(n^p \log^q n)$, $p = 1, 2$, with q independent of d, where $n = N^{1/d}$ is the dimension of the discrete problem in *one* space direction.

AMS Subject Classification: 65F50, 65F30, 46B28, 47A80
Key Words: hierarchical matrices, Kronecker tensor-product, high spatial dimension, Sinc-interpolation, Sinc-quadrature

1 Introduction

In the wide range of applications one requires tractable approximations to certain multi-dimensional nonlocal operators posed in $\mathbb{R}^d$, $d \geq 2$ (see Part I [16] for a more detailed discussion). From the computational point of view the problem is reduced to the efficient representation of the large fully populated matrices in special data-sparse formats.

In [4]-[6], [11] an $\mathcal{H}$-matrix approximation of almost linear complexity to the class of operator-valued functions of an elliptic operator was developed. However, in multidimensional perspective, even approximations with linear complexity $\mathcal{O}(n^d)$ might not be satisfactory. To relax the "curse of dimensionality" we try to represent the corresponding data (matrices and vectors) in the tensor-product form with the overall complexity $\mathcal{O}(dn^p \log^q n)$ with $p, q \geq 1$ independent of d. To optimise the exponent p, we can utilise the hierarchical (i.e. $\mathcal{H}$-matrix) format for each low-dimensional component, which further reduces the cost to $\mathcal{O}(dn \log^q n)$. We recall that the *hierarchical Kronecker tensor-product* (HKT) approximation of a matrix is defined as follows. Given a matrix $A \in \mathbb{C}^{N\times N}$ of dimension $N = n^d$, we try to approximate A by a matrix $A_{(r)}$ of the form

$$A_{(r)} = \sum_{k=1}^{r} V_k^1 \otimes \cdots \otimes V_k^d \;\approx\; A, \tag{1.1}$$

where the V_k^ℓ are $n \times n$-matrices and $\otimes$ denotes the Kronecker product operation. The crucial parameter is r, called the *Kronecker rank* (cf. [1, 17, 20]). Each elementary Kronecker factor V_k^ℓ is supposed to be represented in the $\mathcal{H}$-matrix form.

In the previous paper [7] the HKT approximation is applied to represent the inverse of a discrete elliptic operator in a hypercube $(0,1)^d \in \mathbb{R}^d$, the operator exponential as well as fractional powers of an elliptic operator. The complexity of the HKT approximation can be estimated by $\mathcal{O}(dn \log^q n)$, where q is some fixed constant independent of d. A computational scheme for a low Kronecker-rank approximation to the solution of a tensor system with tensor right-hand side is discussed in [9]. The incremental rank-one approximation

algorithm for orthogonally decomposable tensors is discussed in [21]. Alternative ideas based on sparse grid finite elements/wavelets have been successfully applied [3, 18, 12, 19].

The rest of the paper is organised as follows. In Section 2 we outline the main ideas how to construct the HKT approximations of the operator-valued functions of the discrete elliptic operators with separable coefficients. Section 3 is devoted to the HKT representation of the multi-dimensional integral operators. The inverse A^{-1} of a discrete elliptic operator A is considered in Section 4, while the HKT approximation to the matrix-valued function sign(A) is addressed in Section 5. We remark that sign(A) plays an important role in the density function theory for the electronic structure calculation based on the Hartree-Fock equation [2]. Furthermore, the matrix sign(A) is the main ingredient in the solution process of large scale algebraic matrix Riccati equations in control theory [11].

2 Preliminaries

Boundary/volume integral operators arise in traditional FEM/BEM applications. Multi-dimensional integral operators are, e.g., involved in the classical Schrödinger and Boltzmann equations.

The HKT approximation to integral operators posed in $\mathbb{R}^d$ is well understood, since it can be reduced to a separable approximation of the explicitly given kernel function together with an $\mathcal{H}$-matrix representation of the low-dimensional components (cf. [17] related to the case $d = 2$). In Section 3, we address this topic in the case of rather general shift-invariant kernel functions with $d \geq 2$.

Next we consider the operator-valued functions of an elliptic operator. In the following we use the notation $\mathcal{A}, \mathcal{B}, ...$ for operators and $A, B, ...$ for matrices. The basic assumption on the elliptic operator $\mathcal{A}$ given in the form

$$\mathcal{A} = -\sum_{j=1}^{d} \frac{\partial}{\partial x_j} a_j(x) \frac{\partial}{\partial x_j} + \sum_{j=1}^{d} b_j(x) \frac{\partial}{\partial x_j} + c(x), \quad x = (x_1, ..., x_d) \in (0,1)^d,$$

is the following

$$a_j(x) = a_j(x_j), \quad b_j(x) = b_j(x_j) \quad \text{and} \quad c(x) = c_1(x_1) + ... + c_d(x_d). \tag{2.1}$$

This ensure the existence of a splitting

$$\mathcal{A} = \sum_{j=1}^{d} \mathcal{A}_j, \qquad \mathcal{A}_k \mathcal{A}_m = \mathcal{A}_m \mathcal{A}_k \quad (1 \leq k, m \leq d) \tag{2.2}$$

with mutually commutative differential operators $\mathcal{A}_j$, acting on the variable x_j.

The main idea to prove an HKT approximation to certain operator-valued functions of $\mathcal{A}$ is based on an integral representation to the operator involving the exponential of $\mathcal{A}$. For example, a negative fractional power of $\mathcal{A}$ can be represented by (cf. [7])

$$\mathcal{A}^{-\sigma} = \frac{1}{\Gamma(\sigma)} \int_0^\infty t^{\sigma-1} \mathrm{e}^{-t\mathcal{A}} \mathrm{d}t, \qquad \sigma > 0, \tag{2.3}$$

provided the integral is existing. To derive the tensor-product representation, we consider finite difference (FD) discretisations (three-point stencil in each variable) on a uniform tensor-product grid in $\mathbb{R}^d$ with n degrees of freedom in each spatial direction. The discretisation matrix has the form

$$A = \sum_{j=1}^{d} A_j \qquad \text{with} \quad A, A_j \in \mathbb{R}^{N \times N}, \quad N = n^d,$$

where the mutually commuting matrices A_j have the tensor-product format

$$A_1 = V^1 \otimes I \otimes \ldots \otimes I, \quad A_2 = I \otimes V^2 \otimes \ldots \otimes I, \quad \ldots, \quad A_d = I \otimes \ldots \otimes I \otimes V^d \tag{2.4}$$

with tridiagonal matrices $V^j \in \mathbb{R}^{n \times n}$, $j = 1, \ldots, d$, and the identity $I \in \mathbb{R}^{n \times n}$. For symmetric approximations we require $\Re e\, \sigma(A) \subset [a, \infty)$ with $a > 0$, where $\sigma(A)$ denotes the spectrum of A. In the more general

case we assume that $\|\mathrm{e}^{-tA}\| \le C\mathrm{e}^{-at}$ for $t \ge 0$ with some $a > 0$. Having at hand representation (2.3) written for the discrete operator A, first, we make use of the fundamental property of the exponential function, namely,

$$\exp(-tA) = \prod_{j=1}^{d} \exp(-tA_j), \tag{2.5}$$

for commuting matrices A_j, and second, we apply an exponentially convergent quadrature rule to represent the integral (2.3) by a sum involving only factorised expressions

$$A^{-\sigma} \approx \sum_{k=-M}^{M} c_k t_k^{\sigma-1} \prod_{j=1}^{d} \exp(-t_k A_j), \qquad t_k \in \mathbb{R}.$$

Then we derive the tensor approximation

$$A^{-\sigma} \approx \sum_{k=-M}^{M} c_k\, t_k^{\sigma-1} \bigotimes_{j=1}^{d} \exp(-t_k V^j) =: A_{(r)} \qquad \text{with } r = 2M+1 \ \text{ (cf. (1.1))} \tag{2.6}$$

providing exponential convergence

$$\|A^{-\sigma} - A_{(r)}\| \le C\mathrm{e}^{-s\sqrt{M}}$$

with constants C, s not depending on M (but depending on σ). Finally, given a desired tolerance $\varepsilon = \mathcal{O}(n^{-\beta})$, $\beta > 0$, we apply the $\mathcal{H}$-matrix approximation to each individual exponent $\exp(-t_k V^j)$ manifesting the linear-logarithmic cost $\mathcal{O}(n \log^q n)$ in n (cf. [4, 7] concerning the existence and construction of the corresponding $\mathcal{H}$-matrix approximation). This leads to the approximant $A_{(r)}$ with the complexity bound $\mathcal{O}(dn \log^q n)$. Numerical examples on the HKT approximation to the discrete Laplacian inverse are presented in Part I [16].

Remark 2.1 *Note that with the choice $\mathcal{A} = -\Delta$, the representation (2.6) would be of particular interest in the cases $\sigma = 1/2$ (preconditioning for the Laplace-Beltrami operator $(-\Delta)^{1/2}$, and for the hypersingular integral operator, e.g., in BEM applications), $\sigma = 1$ (inverse Laplacian), and $\sigma = 2$ (inverse biharmonic operator).*

When we study the HKT approximation applied either to a more general class of elliptic operators (say, operators $\mathcal{A}$ with mixed derivatives or indefinite resolvent operators $(zI - \mathcal{A})^{-1}$, $z \in \mathbb{C}$, or to the more general class of operator-valued functions $\mathcal{F}(\mathcal{A})$ (e.g., for $F(\mathcal{A}) = \mathrm{sign}(\mathcal{A})$) then we find that the strong positiveness and commutativity property required above may fail. The goal of this paper is to present more general HKT formats and new approximation techniques which allow to extend the results from [7, 17] to a wider range of interesting applications which include, in particular, functions of operators with mixed derivatives or of indefinite operators. In this paper we focus on the case of matrix-valued functions $F(A)$, where A represents a discrete elliptic operator with separable coefficients (cf. (2.1)) though the results can be applied to a more general class of matrices.

The main ideas behind our approach are related to the following issues:

(**a**) integral representations for the matrices A^{-1}, $(zI - A)^{-1}$, $(zI - A)^{-1} \pm (\bar{z}I - A)^{-1}$ by means of matrix exponentials e^{-tB^2} and e^{-tBB^*} with certain B (see (2.7) and (2.8));

(**b**) separable approximation to matrices of the type $\exp(A_k A_m)$, involved in (2.4) (see Lemma 4.3 below);

(**c**) integral representations for $F(A) = \mathrm{sign}(A)$, which allow exponentially convergent quadratures.

Concerning item (**a**), we propose the representation of A^{-1} by

$$A^{-1} = A \int_0^\infty \mathrm{e}^{-tA^2} \mathrm{d}t \tag{2.7}$$

(see (2.3) with $\sigma = 1$ and with A substituted by A^2), which already allows to treat invertible operators with rather general location of the spectrum (say, the Helmholtz operator $\mathcal{A} = -\Delta - \kappa^2$, $\kappa \in \mathbb{R}$).

In some applications we need the construction for a sum of conjugate resolvents $(zI - A)^{-1} \pm (\bar{z}I - A)^{-1}$, which can be reduced to the case (2.7). For example, the sum of two resolvents can be represented as inverse of the following positive definite matrix,

$$(zI - A)^{-1} + (\bar{z}I - A)^{-1} = 2X(b^2 I + X^2)^{-1}, \qquad z = a + ib,\ X := aI - A.$$

Now $(b^2 I + X^2)^{-1}$ can be approximated in the HKT format using an approximation by exponential sums.

In the general case (e.g., for the single resolvent $(zI - A)^{-1}$), we may use the modification

$$A^{-1} = A^*(AA^*)^{-1} = A^* \int_0^\infty \mathrm{e}^{-tAA^*} \mathrm{d}t. \tag{2.8}$$

The core of our method is based on efficient quadratures for the integrals in (2.7) and (2.8) combined with a data-sparse representation of the matrices $\exp(A_k A_m)$ with commuting tensor products A_k, A_m. Our study of the previous representations applied to discrete elliptic operators of second order then naturally leads to the understanding of how to treat higher order operators, e.g., the biharmonic operator Δ^2.

Remark 2.2 *The representation (2.8) is in particular successful for solving least squares problems given by the normal equation of the form $A^\top Au = A^\top f$.*

The construction of an HKT approximation to the inverse matrix and, in particular, to the resolvent family $(zI - A)^{-1}$, $z \in \mathbb{C}$, allows to approximate a rather general class of matrix-valued analytic functions, which can be represented by the Dunford-Cauchy integral

$$F(A) := \frac{1}{2\pi i} \int_\Gamma F(z)(zI - A)^{-1} \mathrm{d}z, \tag{2.9}$$

where Γ is a curve containing the spectrum of A in its interior. Here $F(z)$ is a real-valued analytic function in the domain containing the interior of Γ. Usually, the curve Γ is chosen symmetrically with respect to the real axis. We assume that the integral in (2.9) is approximated by a proper quadrature formula

$$F_M(A) := \sum_{k=-M}^{M} c_k F(z_k)(z_k I - A)^{-1} \approx F(A) \tag{2.10}$$

(e.g., Sinc quadrature or Gauss-Lobatto quadrature), where the quadrature points are located symmetrically with respect to the real axis, i.e., $z_k = \bar{z}_{-k}$ and, moreover, $c_k = c_{-k}$, $k = 1, \ldots, M$. For generally located $z_k \in \Gamma$, the corresponding term in (2.10) may fail to satisfy the assumption $\Re e\, \lambda > 0$ for all $\lambda \in \sigma(z_k I - A)$. However, if we combine two terms corresponding to the indices k and $-k$, then the result can be represented in the HKT format, and thus, the total sum in (2.10).

The HKT format will be applied to the matrix-valued function $F(A) = \mathrm{sign}(A)$, based on an efficient quadrature for the integral representation

$$\mathrm{sign}(A) = \frac{1}{c_f} \int_{\mathbb{R}_+} \frac{f(tA)}{t} \mathrm{d}t$$

with certain functions f described in §5.1. Note that in the case of Hermitean matrix A, one can apply the polar decomposition $A = \mathrm{sign}(A)\sqrt{AA^*}$, to represent $\mathrm{sign}(A)$ via $A(AA^*)^{-1/2}$.

We recall that the computation of the matrix $\mathrm{sign}(A)$ plays an important role in the density function theory for the electronic structure calculation based on the Hartree-Fock equation [2]. The matrix $\mathrm{sign}(A)$ is the important ingredient in the solution process of large scale algebraic matrix Riccati equations in control theory [11].

3 HKT Approximation of Integral Operators

A principal ingredient in the HKT representation of integral operators in many spatial dimensions is a separable approximation of the multi-variate function representing the kernel of the operator. Given the integral operator $\mathcal{A} : L^2(\Omega) \to L^2(\Omega)$ in $\Omega := [0, 1]^d \in \mathbb{R}^d$, $d \geq 2$,

$$(\mathcal{A}u)(x) := \int_\Omega g(x, y)u(y)\mathrm{d}y, \quad x, y \in \Omega,$$

with some shift-invariant kernel function $g(x,y) = g(|x-y|)$, which, therefore, can be represented in the form

$$g(x,y) = G(\zeta_1,...,\zeta_d) \equiv g\left(\sqrt{\zeta_1^2+...+\zeta_d^2}\right),$$

where $\zeta_\ell = |x_\ell - y_\ell| \in [0,1]$, $\ell = 1,...,d$. With some fixed $0 \le \alpha_0 < 1$, we introduce the auxiliary function

$$F(\zeta_1,...,\zeta_d) := (\zeta_1 \cdots \zeta_{d-1})^{\alpha_0} G(\zeta_1,...,\zeta_d). \tag{3.1}$$

In this section we suppose that a multi-variate function $F : \mathbb{R}^d \to \mathbb{R}$ can be approximated by a separable expansion

$$F_r(\zeta_1,...,\zeta_d) := \sum_{k=1}^{r} \Phi_k^1(\zeta_1)\cdots\Phi_k^d(\zeta_d) \approx F, \tag{3.2}$$

where the set of functions $\{\Phi_k^\ell : \ell = 1,...,d,\ k = 1,...,r\}$ with $\Phi_k^\ell : [0,1] \to \mathbb{R}$ may be fixed or can be chosen adaptively. Various methods for constructing approximations which are exponentially convergent in r are discussed in Part I [16].

We consider a Galerkin scheme with tensor-product test functions

$$\phi^{\mathbf{i}}(x_1,...,x_d) = \phi_1^{i_1}(x_1)\cdot...\cdot\phi_d^{i_d}(x_d), \qquad \mathbf{i} = (i_1,...,i_d),\ i_\ell \in I_n := \{1,...,n\},\ \ell = 1,...,d.$$

Now we approximate the Galerkin stiffness matrix

$$A = \{(\mathcal{A}\phi^{\mathbf{i}}, \phi^{\mathbf{j}})_{L^2}\}_{\mathbf{i},\mathbf{j}\in I_n^d} \in \mathbb{R}^{N\times N}, \quad N = n^d,$$

by a matrix $A_{(r)}$ of the form (1.1), where the V_k^ℓ are $n\times n$ matrices given by

$$V_k^\ell = \left\{\int_0^1 |x_\ell - y_\ell|^{-\alpha_\ell}\, \Phi_k^\ell(|x_\ell - y_\ell|)\phi_\ell^{i_\ell}(x_\ell)\phi_\ell^{j_\ell}(y_\ell)\mathrm{d}x_\ell\mathrm{d}y_\ell\right\}_{i_\ell,j_\ell=1}^{n}, \quad \ell = 1,...,d, \tag{3.3}$$

with $\alpha_\ell = \alpha_0$, $\ell = 1,...,d-1$, and $\alpha_d = 0$ (see (3.1)). We recall the conventional definition of asymptotically smooth functions: a function $g(x,y)$, $x,y \in \mathbb{R}^d$, is called asymptotically smooth if there exists $\gamma \ge 1$, and $p \in \mathbb{R}$ such that for all $x,y \in \mathbb{R}^d$, $x \ne y$, and all multi-indices α,β such that $|\alpha|+|\beta| > 0$ with $|\alpha| = \alpha_1 + ... + \alpha_d$, there holds

$$|\partial_x^\alpha \partial_y^\beta g(x,y)| \le C\alpha!\beta!\gamma^{|\alpha|+|\beta|}|x-y|^{-p-|\alpha|-|\beta|}.$$

The next lemma shows that the error $\|A - A_{(r)}\|$ with respect to usual norms is directly related to the error $\|F - F_r\|_\infty$ of the separable approximation (3.2) of F (see the discussion in [17]). It also specifies the sufficient conditions for $\mathcal{H}$-matrix approximability to the Kronecker factors V_k^ℓ.

Lemma 3.1 *Let (3.2) be valid, then for any $\mathbf{i},\mathbf{j} \in I_n^d$, we have the estimate*

$$|a_{\mathbf{i},\mathbf{j}} - a^r_{\mathbf{i},\mathbf{j}}| \le \|F - F_r\|_\infty \prod_{\ell=1}^{d} \left\| |x_\ell - y_\ell|^{-\alpha_\ell}\, \phi_\ell^{i_\ell}(x_\ell)\phi_\ell^{j_\ell}(y_\ell)\right\|_{L^1([0,1]\times[0,1])}$$

for the components of $A - A_{(r)}$. We assume that the function

$$g_{\ell,k}(u,v) := |u-v|^{-\alpha_\ell}\Phi_k^\ell(|u-v|), \quad (u,v) \in [0,1]^2,$$

is asymptotically smooth for $\ell = 1,...,d,\ k = 1,...,r$. Then, for low-order piecewise polynomial basis functions, V_k^ℓ can be approximated by a rank-m $\mathcal{H}$-matrix $\widetilde{V}_k^\ell$ with an error

$$\|V_k^\ell - \widetilde{V}_k^\ell\| \le C\eta^m \qquad \text{for some} \quad \eta < 1.$$

Proof. By construction we obtain

$$
\begin{aligned}
|a_{\mathbf{i},\mathbf{j}} - a^r_{\mathbf{i},\mathbf{j}}| &= \left| \int_{\Omega\times\Omega} (F - F_r) \left(\prod_{\ell=1}^{d} |x_\ell - y_\ell|^{-\alpha_\ell} \right) \phi^{\mathbf{i}}(x)\phi^{\mathbf{j}}(y) \mathrm{d}x\mathrm{d}y \right| \\
&\le \|F - F_r\|_\infty \left\| \left(\prod_{\ell=1}^{d} |x_\ell - y_\ell|^{-\alpha_\ell} \right) \phi^{\mathbf{i}}(x)\phi^{\mathbf{j}}(y) \right\|_{L^1(\Omega\times\Omega)} \\
&= \|F - F_r\|_\infty \prod_{\ell=1}^{d} \left\| |x_\ell - y_\ell|^{-\alpha_\ell} \phi^{i_\ell}_\ell(x_\ell)\phi^{j_\ell}_\ell(y_\ell) \right\|_{L^1([0,1]\times[0,1])},
\end{aligned}
$$

where the last equation follows by inserting the tensor-product basis and by separating the $2d$-dimensional integral.

To prove the second statement, we note that V^ℓ_k given by (3.3) appears to be the exact Galerkin stiffness matrix for an integral operator with the kernel function $g_{\ell,k}(u,v)$ defined on $[0,1]\times[0,1]$. Since $g_{\ell,k}(u,v)$ is supposed to be asymptotically smooth, the result follows by the conventional theory of the $\mathcal{H}$-matrix approximation (cf. [10], [13]-[15]). ∎

Note that due to Lemma 3.1, $\|A - A_{(r)}\|$ can be easily estimated in, say, the Frobenius, l_2 or l_∞ matrix norms. In particular, we have

$$
\|A - A_{(r)}\|_\infty \le n^d \|F - F_r\|_\infty \prod_{\ell=1}^{d} \left\| |x_\ell - y_\ell|^{-\alpha_\ell} \phi^{i_\ell}_\ell(x_\ell)\phi^{j_\ell}_\ell(y_\ell) \right\|_{L^1([0,1]\times[0,1])}.
$$

Several methods of separable approximations to multi-variate functions are presented in Part I [16]. In the general case, approximability property (3.2) can be validated by using the tensor-product Sinc interpolation. In this case the function $\Phi^\ell_k(|u-v|)$ can be proved to be asymptotically smooth. For the class of kernel functions approximated by the quadrature method or by exponential sums, the factor $\Phi^\ell_k(|u-v|)$ even appears to be globally smooth (indeed, it is the entire function).

Lemma 3.2 *For both, the tensor-product Sinc-interpolation and quadrature approximation methods, the function $g_{\ell,k}(u,v)$ from Lemma 3.1 is asymptotically smooth.*

Proof. In the first case we have

$$
g_{\ell,k}(u,v) = |u-v|^{-\alpha_\ell} S(k,\mathfrak{h})(\phi^{-1}(|u-v|)), \qquad u,v \in [0,1],
$$

where $S(k,\mathfrak{h})$ refers for the k-th Sinc function with step-size $\mathfrak{h}$, and $\phi^{-1}(x) = \mathrm{arsinh}(\mathrm{arcosh}(\frac{1}{x}))$ (cf. [16, §2]). Since the Sinc function $S(k,\mathfrak{h})(x)$, $x \in \mathbb{R}$, is holomorphic in x, and since the factor $|u-v|^{-\alpha_\ell}$ is asymptotically smooth, we conclude that $g_{\ell,k}(u,v)$ is also asymptotically smooth (see also the discussion in [17]).

In the case of a quadrature method, we obtain the entire function

$$
\Phi^\ell_k(|u-v|) = \exp(-t_k|u-v|^2), \quad t_k > 0.
$$

Then the previous argument completes the proof. ∎

Applying Lemmata 3.1 and 3.2 proves the existence of a low Kronecker rank HKT approximation to the class of multi-dimensional integral operators.

In general, given a tolerance $\varepsilon > 0$, we have the bound $r = \mathcal{O}(\left[\log\left(\frac{1}{h}\right)\log\left(\frac{1}{\varepsilon}\right)\log\left(\log\frac{1}{\varepsilon}\right)\right]^{d-1})$, where h is the mesh parameter of the FE discretisation. However, in many practically interesting cases we obtain a dimensionally independent bound $r = \mathcal{O}(\log\left(\frac{1}{h}\right)\log\left(\frac{1}{\varepsilon}\right)\log\left(\log\frac{1}{\varepsilon}\right))$ (see examples in [16]).

4 Approximation to A^{-1} for Indefinite Matrices

The HKT representation to the inverse of a discrete elliptic operator plays a central role when treating the case of general matrix-valued functions $F(A)$, since typically an approximation of the indefinite matrix

resolvent $(zI - A)^{-1}$ is needed (cf. (2.10)). In this case a representation like (2.3) is no longer true. To overcome this difficulty, we approximate the matrix resolvent making use of a sum of resolvents with conjugate parameters z and $\bar{z}$ combined with the integral representations for matrices with positive real part of spectrum. The alternative is a direct representation of $(zI - A)^{-1}$ by (2.8).

Numerical results on the Kronecker-product approximation to the multi-dimensional Laplace operator inverse are presented in [16]. The $\mathcal{H}$-matrix approximation to the resolvent matrix was discussed in [4].

4.1 Using the Representations (2.7) and (2.8)

First, we consider the representation (2.7) provided that the integral exists. For the ease of presentation, we assume that A is real-diagonalisable, i.e., $A = TDT^{-1}$ with D real diagonal, and $\sigma(D^2) \subset [1, R]$ with some $R > 1$. In this case, one can use the efficient quadrature described in Part I [16, Remark 2.2], which yields the representation of the integral term in (2.7) by the exponential sum $F_M(A)$:

$$F(A) := A^{-2} = \int_0^\infty \mathrm{e}^{-tA^2} \mathrm{d}t \approx F_M(A) := \sum_{k=-M}^{M} c_k \mathrm{e}^{-t_k A^2}, \tag{4.1}$$

providing exponentially fast convergence

$$\|F(A) - F_M(A)\| \le C \operatorname{cond}(T) \mathrm{e}^{-\pi M/(\log R \log M)}.$$

Suppose that $\operatorname{cond}(D^2) = n^\beta$, $\beta > 0$, then an approximation error $\varepsilon = n^{-\alpha}$, $\alpha > 0$, can be achieved with $M = \mathcal{O}(\log^q n)$, $q \ge 1$. Figures 4.1 and 4.2 in [16] illustrate the efficiency of the quadrature (4.1) depending on the parameter R bounding the condition: $\sigma(D^2) \subset [1, R]$.

Remark 4.1 *If A is non-diagonalisable, we suppose that $\Re\mathrm{e}\,\sigma(A^2) \in \mathbb{R}_+$. The matrix e^{-tA^2} can be represented by the Dunford-Cauchy integral*

$$\mathrm{e}^{-tA^2} := \frac{1}{2\pi i} \int_\Gamma \mathrm{e}^{-tz} (zI - A^2)^{-1} \mathrm{d}z, \tag{4.2}$$

where Γ is a curve containing the spectrum of A^2 in its interior such that $\Re\mathrm{e}\, z > 0$ for all $z \in \Gamma$. Furthermore, we assume that there is a constant $C_A > 0$ such that

$$\frac{1}{2\pi i} \int_\Gamma \|(zI - A^2)^{-1}\| \, |\mathrm{d}z| \le C_A. \tag{4.3}$$

Denote $\nu = \max\limits_{z \in \Gamma} |\Im\mathrm{m}\, z|$. Now the quadrature error in (4.1) corresponding to [16, Remark 2.2] can be estimated by

$$\|F(A) - F_M(A)\| \le C C_A \, \mathrm{e}^{\pi\nu/2} \, \mathrm{e}^{-\pi M/(\log R \log M)}.$$

We assume that the matrices in (2.4) can be represented in the form

$$A_j = L_j - \frac{\kappa^2}{d} I \qquad \text{with} \quad \Re\mathrm{e}\,\sigma(L_j) \subset [a, \infty),\ a > 0, \text{ and } \kappa^2 \in \mathbb{R}_+$$

(Helmholtz type operators). Due to (2.4), we can write

$$A^2 = \sum_{j=1}^{d} \left(L_j^2 - 2\kappa^2 L_j + \frac{\kappa^4}{d} I \right) + 2 \sum_{1 \le i < j \le d} L_{ij}, \qquad L_{ij} := L_i L_j,$$

which implies (abbreviating $G_j := L_j^2 - 2\kappa^2 L_j + \frac{\kappa^4}{d} I$)

$$F_M(A) = \sum_{k=-M}^{M} S_k \prod_{j=1}^{d} \mathrm{e}^{-t_k G_j} \qquad \text{with} \quad S_k = c_k \prod_{1 \le i < j \le d} \mathrm{e}^{-2t_k L_{ij}}, \tag{4.4}$$

since, by definition, the matrices G_j, L_{ij} $(1 \le i < j \le d)$ commute pairwise. For the finite difference scheme under consideration we can write

$$L_j = I \otimes \ldots \otimes I \otimes B^j \otimes I \otimes \ldots \otimes I \qquad (j \in \{1, \ldots, d\})$$

with $B^j = \mathrm{tridiag}\{a_j, b_j, c_j\} \in \mathbb{R}^{n\times n}$ situated in the j-th position. $I \in \mathbb{R}^{n\times n}$ is the identity. Then, denoting

$$G^j := (B^j)^2 - 2\kappa^2 B^j + \frac{\kappa^4}{d} I \in \mathbb{R}^{n\times n} \tag{4.5}$$

and substituting $B^j - \frac{\kappa^2}{d} I$ (with B^j strongly positive) into the representation (2.4) instead of V^j, we obtain

$$G_1 = G^1 \otimes I \otimes \ldots \otimes I, \quad G_2 = I \otimes G^2 \otimes \ldots \otimes I, \ \ldots, \quad G_d = I \otimes \ldots \otimes I \otimes G^d \tag{4.6}$$

and

$$L_{ij} = I \otimes \ldots \otimes I \otimes B^i \otimes I \otimes \ldots \otimes I \otimes B^j \otimes I \otimes \ldots \otimes I \qquad (1 \le i < j \le d)\,,$$

where the Kronecker factors B^i, B^j correspond to the i-th and j-th position, respectively. In the following, we assume

$$\sigma(L_j) \subset [\lambda_{min}, \lambda_{max}], \quad \lambda_{min}, \lambda_{max} \in \mathbb{R}_{>0} \tag{4.7}$$

(an extension to the case of complex eigenvalues is possible).

Lemma 4.2 *We assume that the matrices in (2.4) have the form $A_j = L_j - \frac{\kappa^2}{d} I$. Rewriting G^j from (4.4) as $G^j = (B^j - \kappa^2 I)^2 + (\frac{1}{d} - 1)\kappa^4 I$, we obtain that the representation*

$$F_M(A) = \sum_{k=-M}^{M} \mathrm{e}^{(d-1)t_k\kappa^4} S_k \bigotimes_{j=1}^{d} \exp(-t_k (B^j - \kappa^2 I)^2) \tag{4.8}$$

provides the error bound $\|F(A) - F_M(A)\| \le C\mathrm{e}^{-\pi M/(\log R \log M)}$. Moreover, $F_M(A)$ can be computed within the tolerance $\varepsilon = n^{-\alpha}$, $\alpha > 0$, with the cost $\mathcal{O}(d^2 n^2 \log R \log^q n)$.

Proof. We start from the representation (4.4). Using (4.6), we immediately obtain

$$\prod_{j=1}^{d} \exp(-t_k G_j) = \bigotimes_{j=1}^{d} \exp(-t_k G^j).$$

The property $(B^j - \kappa^2 I)^2 > 0$ means that the corresponding matrix exponentials can be represented by $\mathcal{H}$-matrices with the cost $\mathcal{O}(n \log^q n)$ (cf. [4, 6]).

In general, the factor S_k in (4.4) cannot be represented exactly in a tensor-product form, hence, the complexity of the product $\prod_{1\le i<j\le d} \exp(-2t_k L_{ij})$ is dominated by the cost for approximating the specific exponential matrices $\exp(-2t_k L_{ij})$ in a data-sparse format. This is possible because, by the definition of L_{ij} and by assumption (4.7), we have $\sigma(L_{ij}) \subset \mathbb{R}_{>0}$. Since L_{ij} operates on an index set isomorphic to $\mathbb{R}^{n^2\times n^2}$, the corresponding cost can be estimated by $\mathcal{O}(d^2 n^2 \log^q n)$ provided that we apply an $\mathcal{H}$-matrix approximation as discussed, e.g., in [7]. ∎

Even under the above conditions, the computational complexity is only quadratic with respect to n rather than exponential in d.

The above representation, due to the square in the exponent, allows to get rid of the restrictions on κ. However, if κ is large, the ansatz (4.8) includes "small" matrix exponentials but multiplied with large coefficients which requires computations with low rounding error.

In the rest of this section, we analyse some cases when the quadratic cost $\mathcal{O}(d^2 n^2 \log^q n)$ in n can be reduced to the linear expense $\mathcal{O}(d^2 n \log^q n)$. Assume that

$$B^j = T \cdot D^j \cdot T^{-1} \in \mathbb{R}^{n\times n}, \qquad j = 1, \ldots, d, \tag{4.9}$$

with real diagonal matrices $D^j = \mathrm{diag}\{\lambda_1^{(j)}, \ldots, \lambda_n^{(j)}\}$ and that T, D^j are numerically available. For example, this is the case if L_j is the finite difference approximation of the one-dimensional Laplacian.

Lemma 4.3 *Let (4.9) hold with real diagonal matrices D^j. Then the matrix $\mathrm{e}^{-2t_k L_{ij}}$ can be represented (up to a tolerance $\varepsilon > 0$) in the Kronecker tensor-product form*

$$\mathrm{e}^{-2t_k L_{ij}} \approx \mathrm{e}^{d-2} \sum_{m=1}^{r_1} I \otimes \ldots \otimes I \otimes \mathcal{D}_{m1}(B^i) \otimes I \otimes \ldots \otimes I \otimes \mathcal{D}_{m2}(B^j) \otimes I \otimes \ldots \otimes I \equiv F_{ij} \tag{4.10}$$

with $r_1 = \mathcal{O}(\log^2(\varepsilon^{-1}))$ and with some explicitly given functions $\mathcal{D}_{m1}$ and $\mathcal{D}_{m2}$ depending on t_k.

Proof. The assertion is a consequence of (4.9) and the corresponding approximation results for the function $\exp(-xy)$, $x, y \geq 0$ (see the example in [16, §7.4]). In fact, (4.9) implies

$$\mathrm{e}^{-2t_k L_{ij}} = T_d \exp\{-2t_k I \otimes \ldots \otimes I \otimes D^i \otimes I \otimes \ldots \otimes I \otimes D^j \otimes I \otimes \ldots \otimes I\} T_d^{-1},$$

where $T_d = T \otimes \ldots \otimes T$. Now the existence of (4.10) is equivalent to the approximability of the exponentials of the Hadamard products

$$\Lambda^{(ij)} = \{\exp(-2t_k \lambda_l^{(i)} \cdot \lambda_m^{(j)})\}_{l,m=1}^n \in \mathbb{R}^{n \times n}, \quad \lambda_l^{(i)} \in \sigma(B^i),\ \lambda_m^{(j)} \in \sigma(B^j),$$

by rank-r_1-matrices with small r_1. In turn, the latter task is reduced to the separable approximation of the generating function $\exp(-2t_k xy)$, $x, y \in [\lambda_{min}, \lambda_{max}] \subset \mathbb{R}_+$, $t_k > 0$, which is accomplished by using a *Sinc* interpolation. For this purpose, we consider the modified function $g(x,y) = \frac{x}{1+x} \exp(-2t_k xy)$, $x \in \mathbb{R}_+$. This function $g(x,y)$ can be approximated by a separable ansatz $g_{M_1}(x,y)$ like in [16, §7.4]. Then we derive

$$|g(x,y) - g_{M_1}(x,y)| \leq (2t_k)^{-1} M_1^{1/2} \mathrm{e}^{-cM_1^{1/2}}.$$

Note that we have $t_k \geq CM^{-1}$ with M corresponding to the exterior sum. Then choosing $r_1 = 2M_1 + 1$, we finally obtain the desired separable representation to the function $\frac{1+x}{x} g(x,y)$, $x \in [\lambda_{min}, \lambda_{max}]$. ∎

Combining (4.10) with (4.8), we arrive at

$$F_M(A) = \sum_{k=-M}^{M} e^{(d-1)t_k \kappa^4} \left\{ \prod_{1 \leq i < j \leq d} F_{ij} \right\} \bigotimes_{j=1}^{d} \mathrm{e}^{-t_k (B^j - \kappa^2 I)^2} \equiv \sum_{k=-M}^{M} \widetilde{S}_k \bigotimes_{j=1}^{d} \mathrm{e}^{-t_k (B^j - \kappa^2 I)^2} \tag{4.11}$$

with F_{ij} from (4.10). This representation is of the generalised format (5.1) with the Kronecker rank $r = 2M + 1$. The complexity of (4.11) can be estimated by $\mathcal{O}(M r_1 d^2 n \log^q n)$ provided that we are able to diagonalise each matrix B^j $(1 \leq j \leq d)$ with the cost $\mathcal{O}(n \log n)$ (e.g., if L_j represents a discrete elliptic operator with constant coefficients). As a consequence, the functions $F_{mp}(B^i)$ can be represented with linear-logarithmic cost in n.

The representation (4.11) can be also transformed to the standard Kronecker-product form since the Kronecker rank of $\widetilde{S}_k$ can be estimated by $r_1^{(d^2-d)/2}$.

4.2 Sum of Conjugate Resolvents

Using quadrature rules like (2.10) it is often possible to choose $c_k = c_{-k}$ and $z_k = \bar{z}_{-k}$. For such cases we propose the following simplifications to represent a sum of conjugate resolvents.

Let $z = a + ib \in \mathbb{C}$ with $a \geq 0$ and a matrix $L = \sum_{j=1}^d L_j$ be given such that

(a1)	$\sigma(L_j) \subset \mathbb{R}_{>0}$	for all $j = 1, ..., d$;
(b1)	$b^2 + \Re e(a - \mu)^2 > c_L > 0$	for all $\mu \in \sigma(L)$.

We are interested in the HKT approximation of the matrices

$$\begin{aligned} \mathcal{G}^+(L) &:= (zI - L)^{-1} + (\bar{z}I - L)^{-1} = 2X(b^2 I + X^2)^{-1}, \\ \mathcal{G}^-(L) &:= (zI - L)^{-1} - (\bar{z}I - L)^{-1} = -2ib(b^2 I + X^2)^{-1} \quad \text{with} \quad X = aI - L. \end{aligned} \tag{4.12}$$

Consider, for example, the matrix $\mathcal{G}^+(L)$. Due to condition (**b1**), one can represent the matrix

$$F(L) := (2X)^{-1}\mathcal{G}^+(L) = (b^2 I + X^2)^{-1}$$

by

$$F(L) = \int_0^\infty \mathrm{e}^{-t(b^2 I + X^2)}\mathrm{d}t \approx F_M(L) := \sum_{k=-M}^{M} c_k \mathrm{e}^{-t_k(b^2 I + X^2)}$$

(cf. (4.1)). Using the same notation as in §4.1, we substitute $a = \kappa^2$, $A_j = L_j - \frac{a}{d} I$, to obtain

$$b^2 I + X^2 = \sum_{j=1}^{d} \left(L_j^2 - 2aL_j + \frac{a^2 + b^2}{d} I \right) + 2 \sum_{1 \le i < j \le d} L_{ij}, \qquad L_{ij} := L_i L_j.$$

Taking into account condition (**a1**), we are able to apply Lemma 4.2 which, in our particular case, leads to the representation of the form (5.1),

$$F_M(L) = \sum_{m=-M}^{M} \mathrm{e}^{t_m(a^2 - \frac{a^2+b^2}{d})} S_m \bigotimes_{j=1}^{d} \exp(-t_m(B^j - aI)^2) \approx F(L), \tag{4.13}$$

with S_k given by (4.4). Opposite to the discussion following Lemma 4.2, the expression (4.13) does not indicate any numerical instabilities because of the following remark.

Remark 4.4 *The Kronecker sum (4.13) includes only coefficients* $\mathrm{e}^{t_m(a^2 - \frac{a^2+b^2}{d})} S_m$*, which can be bounded by* $\mathcal{O}(\varepsilon^{-q})$ *with* $q = \mathcal{O}(1)$*, in the* L_∞*-norm, which do not cause numerical instabilities. In fact, by construction we easily obtain* $\max\limits_m \{t_m\} \le CM = \mathcal{O}(|\log \varepsilon|)$*, where* ε *is the required accuracy (cf. Part I [16]). Moreover, by the same reasons, we can show that the quadrature points in (2.10) applied to the Dunford-Cauchy integral also satisfy* $\max\limits_{-M \le k \le M} \sqrt{a_k^2 + b_k^2} \le CM$*, where* $z_k = a_k + ib_k$*.*

Representation (4.13) not only proves the existence of an HKT approximation to the considered class of matrix-valued functions but also provides a constructive algorithm for computing such an approximation in real arithmetic. Similarly to (4.11), the corresponding cost is dominated by $\mathcal{O}(Mrd^2 n \log^q n)$.

If we are interested to approximate the individual resolvents $(zI - L)^{-1}$, $z \in \mathbb{C}$, we can apply the representation (2.8). The corresponding construction is completely similar to the previous case. The only difference is in the usage of complex arithmetics to multiply A^* with a real valued integral representing the matrix $F(L)$ analysed above. The same recipe can be used to approximate A^{-1} in the case of an invertible matrix with a rather general location of the spectrum $\sigma(A)$.

4.3 Approximation to a Class of Analytic Functions $F(A)$

Assume that a given matrix-valued analytic function $F(A)$ allows the Dunford-Cauchy representation. Assume that we are given a quadrature rule (2.10) with symmetric coefficients and quadrature points (i.e., $c_k = c_{-k}$, $z_k = \bar{z}_{-k}$, $k = 1, \ldots, M$) that converges exponentially in M, i.e.,

$$||F(A) - F_M(A)|| \le C\mathrm{e}^{-cM^\alpha} \quad \text{for some } \alpha > 0.$$

For our particular quadratures we have $\alpha \ge 1/2$. Now, we apply the HKT approximation to each couple of conjugate resolvents (cf. §4.2) to obtain a method of complexity $\mathcal{O}(Mrd^2 n \log^q n)$ provided that we are able to diagonalise each matrix B^j.

5 HKT Approximation to $\mathrm{sign}(A)$

In certain cases, each term in (1.1) may be amplified by an extra factor $S_k \in \mathbb{R}^{N \times N}$. For this purpose, we introduce the generalised tensor-product matrix format

$$A_{(r)} = \sum_{k=1}^{r} S_k \cdot \left(V_k^1 \times \cdots \times V_k^d \right) \;\approx\; A \tag{5.1}$$

with a matrix S_k having a special data-sparse representation of complexity $\mathcal{O}(n^p \log^q n)$ with $p \le 2$ (cf.(4.11)). In this section, the format (5.1) will be applied to the matrix-valued function $F(A) = \operatorname{sign}(A)$. In this case all factors S_k can be represented in the rank-r_1 KHT format and thus the whole sum can be converted into the KHT format with Kronecker rank rr_1.

The matrix sign function of $A \in \mathbb{R}^{N\times N}$ is defined by

$$\operatorname{sign}(A) := \frac{1}{\pi i}\int_{\Gamma_+} (zI - A)^{-1}\mathrm{d}z - I \tag{5.2}$$

with Γ_+ being any simply closed curve in the complex plane whose interior contains all eigenvalues of A with positive real part. The first approach to approximate $F(A) = \operatorname{sign}(A)$ in the HKT format is based on the efficient quadrature (2.10) (cf. [6] concerning the existence), then the corresponding Kronecker rank is $r = 2M + 1$.

A second method to construct the HKT representation to $\operatorname{sign}(A)$ is based on the matrix version of the integral representation

$$\operatorname{sign}(a) = \frac{1}{c_f}\int_0^\infty \frac{f(ta)}{t}\mathrm{d}t, \quad c_f > 0, \tag{5.3}$$

where $a \in \mathbb{C}$, $\Re e\, a \neq 0$, with $f : \mathbb{C} \to \mathbb{C}$ satisfying certain assumptions discussed in §5.1.

To derive efficient representations (1.1) or (5.1), the function f has to be chosen in such a way that the integral (5.3) allows exponentially convergent quadrature and, moreover, f facilitates a good "separability property". We show that the following examples are satisfactory choices:

$$f_1(t) := t\exp(-t^2), \tag{5.4a}$$

$$f_2(t) := \frac{t}{1+\alpha t^2}, \quad \alpha > 0, \tag{5.4b}$$

$$f_{3,n}(t) := \frac{j_n(t)}{t^{n-1}}, \quad n = 1, 2, \ldots, \tag{5.4c}$$

where $j_n(t)$ are the spherical Bessel functions of the first kind (cf. [8]). In particular, we have

$$j_0(t) = \frac{\sin(t)}{t}, \quad j_1(t) = \frac{\sin(t) - t\,\cos(t)}{t^2}, \quad j_2(t) = \left(\frac{3}{t^3} - \frac{1}{t}\right)\sin(t) - \frac{3}{t^2}\cos(t).$$

Note that the integral representations (2.7) and (5.5) (the latter with the choice $f(t) = t\exp(-t^2)$) applied to the matrices A^{-1} and $\operatorname{sign}(A)$, respectively, lead to rather similar expressions. Representation (5.3) may have different advantages and limitations depending on the particular choice of the function f and the properties of A (selfadjoint, diagonalisable or rather general).

In §5.1, we analyse certain representations involving the integrands (5.4a-c) in more detail. In particular, we show that the generating functions $f_1(t)$, $f_2(t)$ can be applied to a rather general class of matrices provided that $\Re e\,\sigma(A^2) \subset \mathbb{R}_{>0}$. The corresponding complexity is proved to be $\mathcal{O}(d^2 n^2 \log^q n)$. In the case of real-diagonalisable matrices one can use generating functions $f_{3,n}(t)$ expecting the complexity $\mathcal{O}(dn \log^q n)$.

5.1 Using the Representation (5.3)

We consider two classes of matrices:

Case (A) Let A be real-diagonalisable, i.e., $A = T\,D\,T^{-1}$, where D is real diagonal. Now the function $f : \mathbb{R} \to \mathbb{R}$ is supposed to have the properties

(A1) $f(t) = -f(-t)$, $t \in \mathbb{R}$,

(A2) $c_f := \int_0^\infty \frac{f(t)}{t}\mathrm{d}t \in (0, \infty)$ exists as an improper integral.

Case (B) We assume that $\sigma(A) = \sigma_+(A) \cup \sigma_-(A)$, where

$$\sigma_+(A) := \{\lambda \in \sigma(A) : \Re e\,\lambda > 0\}, \qquad \sigma_-(A) := \{\lambda \in \sigma(A) : \Re e\,\lambda < 0\}.$$

The function $f : \mathbb{C} \to \mathbb{C}$ is supposed to have the properties

(B1) $f(t) = -f(-t), \quad t \in \mathbb{R}$.

(B2) The function $f : \mathbb{C} \to \mathbb{C}$ is analytic in the domain $\Omega = \Omega_+ \cup \Omega_-$ with boundary $\Gamma = \Gamma_+ \cup \Gamma_-$, which is the union of two closed simply connected curves Γ_+ and Γ_-, each of which contains the respective part of the spectrum $\sigma_\pm$. Moreover,

$$|f(z)| \le C(1+|z|)^{-1} \qquad \text{for all } z \in \Omega_\theta := \{z : |\arg(z)| \le \theta < \frac{\pi}{2}\} \quad \text{with } \Omega \subset \Omega_\theta.$$

(B3) For any $z = re^{i\theta} \in \Gamma$, we have $c_f := \int_0^\infty \frac{f(u)}{u} \mathrm{d}u \in (0,\infty)$, where c_f does not depend on z, with the integration path running along the ray $\{u : u = \rho e^{i\theta}, \rho \in [0,\infty)\}$.

Note that in both cases, f is thought to allow an efficient quadrature for (5.3).
Based on formula (5.3), we derive the integral representation to the matrix $\mathrm{sign}(A)$.

Lemma 5.1 *Let A be a square matrix A such that $0 \notin \Re e\, \sigma(A)$. Let the function f satisfy the Assumptions (A1)-(A2) or (B1)-(B3) in the respective Cases (A) or (B). Then we have*

$$\mathrm{sign}(A) = \frac{1}{c_f} \int_{\mathbb{R}_+} \frac{f(tA)}{t} \mathrm{d}t. \tag{5.5}$$

Proof. First we note that for $a \in \mathbb{R} \setminus \{0\}$, the assumptions (A1)-(A2) imply (5.3), while for $a \in \mathbb{C}$ with $\Re e\, a \neq 0$, (B1)-(B3) also yield (5.3).
In Case (A) we have $A = T\,D\,T^{-1}$, so that

$$f(tA) = T\, f(tD)\, T^{-1}. \tag{5.6}$$

Moreover, $\mathrm{sign}(A) = T\ \mathrm{sign}(D)\, T^{-1}$ holds and (5.3) implies the desired relation:

$$\frac{1}{c_f} \int_{\mathbb{R}_+} \frac{f(tA)}{t} \mathrm{d}t \ = \ T \left(\frac{1}{c_f} \int_{\mathbb{R}_+} \frac{f(tD)}{t} \mathrm{d}t \right) T^{-1} = \ T\ \mathrm{sign}(D)\, T^{-1} = \mathrm{sign}(A).$$

In Case (B), the analytic function $f : \mathbb{C} \to \mathbb{C}$ generates the family of matrix-valued functions $f(tA)$, $t \ge 0$, which can be represented by the Dunford-Cauchy integral

$$f(tA) = \frac{1}{2\pi i} \int_\Gamma f(tz)(zI - A)^{-1} \mathrm{d}z, \tag{5.7}$$

where $\Gamma = \Gamma_+ \cup \Gamma_-$ is the union of two closed simply connected curves Γ_+ and Γ_-, each of which contains the respective part $\sigma_\pm$ of the spectrum (cf. assumption (B2)). Note that $\Gamma_\pm$ can be chosen in such a way that with some positive constant μ, the relation $|\Re e\, z| > \mu > 0$ holds for $z \in \Gamma_\pm$. Now due to assumption (B3), we obtain

$$\|f(tA)\| \le c \int_\Gamma |f(tz)|\, \|(zI - A)^{-1}\|\, |\mathrm{d}z| \le \frac{c}{1+t} \int_\Gamma \frac{1}{1+|z|} \|(zI - A)^{-1}\|\, |\mathrm{d}z|,$$

which proves the existence of the integral in (5.5). Let us introduce the integrals

$$B_+ = \frac{1}{\pi i} \int_{\Gamma_+} (zI - A)^{-1} \mathrm{d}z, \qquad B_- = \frac{1}{\pi i} \int_{\Gamma_-} (zI - A)^{-1} \mathrm{d}z.$$

By definition of Γ_+ and Γ_- we have

$$\frac{1}{2}(B_+ + B_-) = I, \quad \text{and thus } B_- = 2I - B_+. \tag{5.8}$$

We substitute the Dunford-Cauchy integral (5.7) into (5.5) and use (5.8) to derive

$$\begin{aligned}\frac{1}{c_f}\int_{\mathbb{R}_+}\frac{f(tA)}{t}\mathrm{d}t &= \frac{1}{2\pi i}\int_\Gamma\big[\frac{1}{c_f}\int_{\mathbb{R}_+}\frac{f(tz)}{t}\mathrm{d}t\big](zI-A)^{-1}\mathrm{d}z\\ &= \frac{1}{2\pi i}\int_{\Gamma_+\cup\Gamma_-}\mathrm{sign}(z)(zI-A)^{-1}\mathrm{d}z\\ &= \frac{1}{2}(B_+-B_-) = B_+ - I\\ &= \mathrm{sign}(A),\end{aligned}$$

which completes the proof. ∎

5.2 Analysis for the Integrands (5.4a) and (5.4b)

In both Cases (A) and (B), we derive an efficient quadrature for the choice $f = f_1(t)$ in (5.3). Let (2.2) be valid and let $\Re e\,\lambda \in [1, R]$ for all $\lambda \in \sigma(A^2)$. We also assume (4.3) to be valid. We approximate the integral (5.5) with $f = f_1$ by applying an exponentially convergent quadrature rule to the integral

$$\int_0^\infty \exp(-t^2A^2)\mathrm{d}t = \frac{1}{2}\int_{\mathbb{R}}\exp(-t^2A^2)\mathrm{d}t$$

appearing in

$$F(A) := A^{-1}\,\mathrm{sign}(A) = \frac{1}{\sqrt{\pi}}\int_{\mathbb{R}}\mathrm{e}^{-t^2A^2}\mathrm{d}t \approx \sum_{k=-M}^{M} c_k\mathrm{e}^{-t_k^2A^2} =: F_M(A) \tag{5.9}$$

(we set $c_f = \sqrt{\pi}$) with c_k, t_k given in [16]. Due to (2.2), we can use the same techniques as in §4.1 to represent each individual exponent in (5.9) in tensor-product form, which leads to the cost $\mathcal{O}(Md^2n^2\log^q n)$ with $M = \mathcal{O}(\log(R)\log\varepsilon^{-1} + C_A + \pi\nu/2)$ or with $M = \mathcal{O}(\log^2\varepsilon^{-1} + C_A + \pi\nu/2)$ depending on the relation between ε^{-1} and R.

Note that in the case $f = f_2(t)$, we have

$$\mathrm{sign}(A) = \frac{A}{c_f(\alpha)}\int_0^\infty (I + \alpha t^2A^2)^{-1}\mathrm{d}t,$$

which is similar to the familiar Robert's integral representation

$$\mathrm{sign}(A) = \frac{2A}{\pi}\int_0^\infty (t^2I + A^2)^{-1}\mathrm{d}t.$$

This case can be reduced to the analysis of the matrix $\mathcal{G}$ in (4.12), therefore, all the results in §4.2 can be applied as well.

5.3 Construction in Case (5.4c)

In Case (A), we may consider as well the generating functions $f = f_{3,n}$ from (5.4c). The spherical Bessel functions $j_n(z)$ (cf. [8]) have the asymptotical property

$$z^{-n}j_n(z) \to \frac{1}{1\cdot 3\cdot 5\ldots(2n-1)} \quad \text{as } z\to 0 \qquad (n = 0, 1, 2, \ldots).$$

We also use the integral representation

$$j_n(z) = \frac{z^n}{2^{n+1}n!}\int_0^\pi \cos(z\cos\theta)\sin^{2n+1}\theta\,\mathrm{d}\theta \qquad (n = 0, 1, 2, \ldots). \tag{5.10}$$

Since the matrix A is diagonalisable, the error analysis of the quadrature rule is reduced to the scalar case.

Let us construct an exponentially convergent quadrature for (5.3) with $f = f_{3,n}$ and with $a \in \mathbb{R}$. In general, one can expect $a \in [1, \Lambda]$ with $1 \ll \Lambda$, so we deal with the integration of a highly oscillatory function

with a smooth weight. For instance, for $n = 1$ the corresponding integrand takes the form $f_{3,1}(at)/t = \frac{\sin(at) - t\cos(at)}{t^3}$, where $a > 0$ is a large parameter. We recall that

$$j_n(z) = g_n(z)\sin z + (-1)^{n+1} g_{-n-1}(z)\cos z \tag{5.11}$$

(cf. [8]), where $g_0(z) = z^{-1}$, $g_1(z) = z^{-2}$, $g_{n-1}(z) + g_{n+1}(z) = (2n+1)z^{-1}g_n(z)$ for $n \in \mathbb{Z}$.

(5.11) yields the estimate $|j_n(t)| \le C/t$, $t \to +\infty$, hence we have

$$\frac{f_{3,n}(at)}{t} = \frac{j_n(at)}{a^{n-1}t^n} \le \frac{C}{at^2}, \qquad t \to \infty. \tag{5.12}$$

The latter implies

$$\left| \int_R^\infty \frac{f_{3,n}(at)}{t} \mathrm{d}t \right| \le \frac{C}{aR}, \qquad R > 0.$$

Moreover, due to (5.10) , $\frac{f_{3,n}(az)}{z}$ is holomorphic at $z = 0$ (in fact, it is an entire function).

Now, given a tolerance $\varepsilon > 0$, we choose $R > 0$ such that $R^{-1} = a\varepsilon$, i.e., $R = (a\varepsilon)^{-1}$, and then construct a quadrature on the finite interval $[0, R]$. Recall that $a^{-1} \in [\Lambda^{-1}, 1]$. We can assume without loss of generality that $\Lambda = 2^{K_0}$ with some $K_0 \in \mathbb{N}$, so that $a^{-1} \in [2^{-K_0}, 1]$.

Again, we split $[0, R]$ into the two parts $[0, 2^{-K_0}]$ and $\omega := [2^{-K_0}, R]$. We choose the number $K_1 \in \mathbb{N}$ such that $K_1 = |\log\varepsilon| + K_0$, provided that $\min_{\lambda\in\sigma_+(A)} \lambda = \mathcal{O}(1)$. Without loss of generality we further assume that $R = 2^{K_1}$. We now decompose the integration interval $\omega = \bigcup_{k=-K_0}^{K_1} [b_k, b_{k+1}]$ by the points $b_k = 2^k$, $k = -K_0, \dots, 0, \dots, K_1$.

Since $g_k(z)$ from (5.11) is a polynomial in z^{-1}, it can be approximated on each interval $\delta_k = [b_k, b_{k+1}]$ by a polynomial $\mathcal{P}_{p,k}$ of degree p such that

$$\max_{t\in\delta_k} |g_k(t) - \mathcal{P}_{p,k}(t)| \le C\mathrm{e}^{-cp}. \tag{5.13}$$

Next we use the integrals

$$\int_0^x t^m \sin(at)dt = -\sum_{k=0}^m k!\tbinom{m}{k} \frac{x^{m-k}}{a^{k+1}} \cos\left(ax + \frac{1}{2}k\pi\right),$$
$$\int_0^x t^m \cos(at)dt = \sum_{k=0}^m k!\,\tbinom{m}{k}\, \frac{x^{m-k}}{a^{k+1}}\, \sin\left(ax + \frac{1}{2}k\pi\right)$$

(cf. [8]), to obtain the following approximation on the interval ω:

$$\frac{1}{c_f}\int_\omega \frac{f_{3,n}(at)}{t}\mathrm{d}t \simeq \sum_{k=-K_0}^{K_1}\sum_{\ell=0}^{p} \left[\gamma_{k\ell}\,\sin(as_{k\ell}) + \mu_{k\ell}\,\cos(ac_{k\ell})\right],$$

which provides an exponential convergence of the order $\mathcal{O}(\mathrm{e}^{-cp})$.

Due to (5.10), the integrand $\frac{f_{3,n}(az)}{z}$ is an entire function and, in particular, holomorphic in the Bernstein ellipse $\mathcal{E}_\rho$ with $\rho > 1/(2a)$, corresponding to the interval $[0, a^{-1}]$ (cf. [15]). Furthermore, $\max_{z\in\mathcal{E}_\rho} \left|\frac{f_{3,n}(az)}{z}\right|$ can be estimated by a constant not depending on a. Therefore, the Gauss quadrature on $[0, \Lambda^{-1}]$ is exponentially convergent. This yields the approximation

$$\mathrm{sign}(\lambda) \sim \mathrm{sign}_M(\lambda) := \sum_{k=1}^M a_k \sin(s_k\lambda) + b_k\cos(c_k\lambda), \tag{5.14}$$

such that for $\lambda \in [1, \Lambda]$ there holds

$$|\,\mathrm{sign}(\lambda) - \mathrm{sign}_M(\lambda)| \le C(K_0 + K_1)\,\mathrm{e}^{-cp}$$

with

$$K_1 = |\log\varepsilon|, \quad K_0 = \log(\mathrm{cond}(A)), \quad M := (K_0 + K_1)\,p. \tag{5.15}$$

Lemma 5.2 *Let A be symmetric with $\min_{\lambda\in\sigma_+(A)}\lambda = \mathcal{O}(1)$. Then, given $\varepsilon > 0$, the quadrature points and weights from (5.14) and (5.15) fulfil*

$$\left\| \frac{1}{c_f}\int_0^\infty \frac{f_{3,n}(tA)}{t}\mathrm{d}t - \sum_{k=1}^{M}\left[a_k(A^{-1})\sin(s_kA) + b_k(A^{-1})\cos(c_kA)\right]\right\|_2 \le C\,\mathrm{cond}(T)(K_0+K_1)\mathrm{e}^{-cp}, \quad (5.16)$$

where p is defined by the choice of the polynomial $\mathcal{P}_{p,k}$ in (5.13), M, K_0, K_1 are explained in (5.15) and $a_k(A^{-1}), b_k(A^{-1})$ are polynomials of A^{-1}.

Proof. Since $A = TDT^{-1}$, we use the representation (5.6), where D has real entries, and derive

$$\begin{aligned}
&\left\| \frac{1}{c_f}\int_{\mathbb{R}_+} \frac{f_{3,n}(tA)}{t}\,\mathrm{d}t - \sum_{k=1}^{M}\left[a_k\,\sin(s_kA) + b_k\,\cos(c_kA)\right]\right\|_2 \\
&= \left\| T\left(\frac{1}{c_f}\int_{\mathbb{R}_+} \frac{f_{3,n}(tD)}{t}\,\mathrm{d}t - \sum_{k=1}^{M}\left[a_k\,\sin(s_kD) + b_k\,\cos(c_kD)\right]\right)T^{-1}\right\|_2 \\
&\le \mathrm{cond}(T)\max_{\lambda\in\sigma_+(A)}\left|\frac{1}{c_f}\int_{\mathbb{R}_+} \frac{f_{3,n}(t\lambda)}{t}\,\mathrm{d}t - \sum_{k=1}^{M}\left[a_k\,\sin(s_k\lambda) + b_k\,\cos(c_k\lambda)\right]\right| \\
&\le C\;\mathrm{cond}(T)\,[K_0+K_1]\,\mathrm{e}^{-cp}.
\end{aligned}$$

Since $M = p(K_0+K_1)$ (cf. (5.15)) the proof is complete. ∎

Note that the simplest possible approximation can be constructed with the choice $f = f_{3,1}$.

To complete this section, we derive tensor-product representations of the matrices $\sin(s_kA)$ and $\cos(c_kA)$ involved in (5.16). For this purpose, we apply the following proposition which can be proved by induction. In the case $d = 2$, the assertion (5.17) is trivial.

Proposition 5.3 ([1]) *Let $d \ge 2$. The trigonometric identity*

$$\sin\left(\sum_{j=1}^{d} x_j\right) = \sum_{j=1}^{d}\sin(x_j)\prod_{k\in\{1,\dots,d\}\setminus\{j\}}\frac{\sin(x_k+\alpha_k-\alpha_j)}{\sin(\alpha_k-\alpha_j)} \quad (5.17)$$

holds for all choices of $\{\alpha_1,\dots,\alpha_d\}$ such that $\sin(\alpha_k-\alpha_j)\neq 0$ for all $j\neq k$.

The following statement extends the trigonometric identity (5.17) to the case of matrix-valued functions $\sin(A)$ and $\cos(A)$.

Corollary 5.4 *Let $A = \sum_{j=1}^{d} A_j \in \mathbb{R}^{N\times N}$ with matrices A_j of the form (2.4), where $V^j \in \mathbb{R}^{n\times n}$ $(j = 1,\dots,d)$ and $N = n^d$. Suppose that $\{\alpha_1,\dots,\alpha_d\}\subset\mathbb{R}$ are chosen in such a way that the representation (5.17) is valid. Then the following tensor-product representation with exactly d terms*

$$\sin(A) = \sum_{j=1}^{d}\bigotimes_{k=1}^{d}\beta_{kj}\sin(V^j + (\alpha_k-\alpha_j)I), \qquad \beta_{kj} = \begin{cases} 1/\left(\sin(\alpha_k-\alpha_j)\right) & k\neq j, \\ 1 & k = j, \end{cases} \quad (5.18)$$

holds. A similar representation exists for the matrix $\cos(A)$.

To guarantee the stability of representation (5.18) we have to control the condition $|\alpha_k-\alpha_j-m\pi| > \delta > 0$ for $m\in\mathbb{Z}$, $k\neq j$.

Lemma 5.2 and Corollary 5.4 lead to the desired Kronecker tensor-product representation of the matrix $\mathrm{sign}(A)$ having the complexity $\mathcal{O}(dMn\log^q n)$ provided that each V^j $(j = 1,\dots,d)$ can be diagonalised with the cost $\mathcal{O}(n\log^q n)$.

5.4 Dunford-Cauchy Integral (5.2) Revisited

If some of the assumptions in **Cases (A), (B)** are not satisfied, one can apply the integral representation (5.2). The exponentially convergent quadrature

$$\operatorname{sign}(A) \approx \sum_{k=1}^{r} c_k (z_k I - A)^{-1} - I, \qquad r = \mathcal{O}\left(\log^2 \varepsilon + \log^2 \operatorname{cond}(A)\right),$$

for the integral (5.2) provides a direct approximation of $F(A) = \operatorname{sign}(A)$ by a sum of matrix resolvents (cf. [6]). The quadrature points and weights can be chosen symmetrically. Using the results in §4, we are led to the overall cost $\mathcal{O}(rd^2n^2 \log^q n)$ in the multi-dimensional case. Again, the complexity is quadratic in d and n.

Acknowledgement. Discussions with Prof. I. Gavrilyuk (Berufsakademie Eisenach, Germany) are gratefully acknowledged.

References

[1] G. Beylkin and M.J. Mohlenkamp: *Numerical operator calculus in higher dimensions,*. Proc. Natl. Acad. Sci. USA, **99** (2002), 10246-10251.

[2] H.-J. Flad, W. Hackbusch, B.N. Khoromskij, and R. Schneider: *Concept of data-sparse tensor-product approximation in many-particle models* (in preparation).

[3] H.-J. Flad, W. Hackbusch, D. Kolb, and R. Schneider: *Wavelet approximation of correlated wavefunctions. I. Basics*, J. Chem. Phys. **116**, (2002), 9641-9657.

[4] I.P. Gavrilyuk, W. Hackbusch, and B.N. Khoromskij: *$\mathcal{H}$-matrix approximation for the operator exponential with applications.* Numer. Math. **92** (2002), 83-111.

[5] I.P. Gavrilyuk, W. Hackbusch, and B.N. Khoromskij: *Data-sparse approximation to operator-valued functions of elliptic operators.* Math. Comp. **73** (2004), 1297–1324.

[6] I.P. Gavrilyuk, W. Hackbusch, and B.N. Khoromskij: *Data-sparse approximation to a class of operator-valued functions.* Math. Comp. **74** (2005), 681-708.

[7] I. P. Gavrilyuk, W. Hackbusch, and B. N. Khoromskij: *Tensor-product approximation to elliptic and parabolic solution operators in higher dimensions.* Preprint 83, Max-Planck-Institut für Mathematik in den Naturwissenschaften, Leipzig 2003; Computing (to appear).

[8] I.S. Gradshteyn and I.M. Ryzhik: *Table of Integrals, Series and Products* (6th edition), Academic Press, San Diego, 2000.

[9] L. Grasedyck: *Existence and computation of a low Kronecker-rank approximation to the solution of a tensor system with tensor right-hand side.* Computing **72** (2004), 247-265.

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

[11] L. Grasedyck, W. Hackbusch, and B.N. Khoromskij: *Solution of large scale algebraic matrix Riccati equations by use of hierarchical matrices.* Computing **70** (2003), 121-165.

[12] M. Griebel and S. Knapek: *Optimized tensor-product approximation spaces.* Constr. Approx. **16** (2000), 303-332.

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

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

[15] W. Hackbusch and B.N. Khoromskij: *Towards $\mathcal{H}$-matrix approximation of linear complexity*. Operator Theory: Advances and Applications **121**, Birkhäuser-Verlag, Basel, 2001, pp. 194-220.

[16] W. Hackbusch and B.N. Khoromskij: *Low-rank Kronecker product approximation to multi-dimensional nonlocal operators. Part I. Separable approximation of multi-variate functions.* Preprint 29, Max-Planck-Institut für Mathematik in den Naturwissenschaften, Leipzig 2005.

[17] W. Hackbusch, B.N. Khoromskij, and E. Tyrtyshnikov: *Hierarchical Kronecker tensor-product approximation*. J. Numer. Math. **13** (2005), 207-243.

[18] H. Luo, D. Kolb, H.-J. Flad, W. Hackbusch, and T. Koprucki: *Wavelet approximation of correlated wavefunctions. II. Hyperbolic wavelets and adaptive approximation schemes*, J. Chem. Phys. **117**, (2002), 3625-3638.

[19] V.N. Temlyakov: *Approximation of functions with bounded mixed derivative*. Proc. Steklov Inst. Math. **178**, No.1 (1989), 275-293.

[20] E.E. Tyrtyshnikov: *Tensor approximations of matrices generated by asymptotically smooth functions.* (Russian) Mat. Sb. **194** (2003), no. 6, 147–160; translation in Sb. Math. **194** (2003), no. 5-6, 941–954.

[21] T. Zhang and G.H. Golub: Rank-one approximation to high order tensors. SIAM J. Matrix Anal. Appl. **23** (2001), 534-550.