BIT Numerical Mathematics (2023) 63:30
https://doi.org/10.1007/s10543-023-00974-y

BIT

Low-rank tensor structure preservation in fractional operators by means of exponential sums

Angelo Casulli[1] · **Leonardo Robol**[2]

Received: 11 August 2022 / Accepted: 18 April 2023 / Published online: 11 May 2023

Abstract
The use of fractional differential equations is a key tool in modeling non-local phenomena. Often, an efficient scheme for solving a linear system involving the discretization of a fractional operator is computing inverse fractional powers of the standard discretized Laplace operator. In this work, an exponential sum approximation for such fractional powers is derived. It is accurate over all positive real numbers larger than one, and allows to efficiently approximate the action of such operators on tensors stored in a variety of low-rank formats (CP, TT, Tucker). The results are relevant from a practical and theoretical perspective, as they predict the low-rank approximability of the solutions of these linear systems in low-rank tensor formats.

Keywords Exponential sums · Fractional powers · Matrix functions · Kronecker sums

Mathematics Subject Classification 65F60 · 15A16 · 30E10 · 15A69

1 Introduction

We are concerned with computing the solution of a linear system $\mathcal{A}^{\alpha} x = c$ for $0 < \alpha < 1$, where $\mathcal{A}$ is a *Kronecker sum*:

Communicated by Michiel E. Hochstenbach.

✉ Leonardo Robol
leonardo.robol@unipi.it

Angelo Casulli
angelo.casulli@sns.it

[1] Scuola Normale Superiore, Pisa, Italy

[2] Department of Mathematics, University of Pisa, Pisa, Italy

$$\mathcal{A} := \bigoplus_{i=1}^{d} A_i = \underbrace{A_1 \otimes I \otimes \ldots \otimes I}_{d \text{ terms}} + I \otimes A_2 \otimes \ldots \otimes I + \ldots + I \otimes \ldots \otimes I \otimes A_d, \qquad (1)$$

for $i = 1, \ldots, d$. This problem arises naturally when solving fractional PDEs on tensorized domains [20, 25, 37] such as approximating the steady-state behavior of the initial value problem

$$\frac{\partial u}{\partial t} = -(-\Delta)^{\alpha} u + f, \qquad \Delta u := \frac{\partial^2}{\partial x_1^2} + \cdots + \frac{\partial^2}{\partial x_d^2}, \qquad u(t, x_1, \ldots, x_d) : [0,1]^d \to \mathbb{R}.$$

A common approach to approximate the differential operator $-(-\Delta)^{\alpha}$ is to discretize the Laplace operator Δ, and then raise the discrete operator to the αth power (adjusting the sign to make it positive definite). This yields a discretization of the fractional Laplacian [20], and whenever the domain has a tensor structure (as in the case above where $\Omega = [0,1]^d$) the discrete operator is a power of a Kronecker sum as in (1). Such structure is directly available if the problem is discretized through finite differences, and can be recovered with finite elements up to inverting mass matrices.

When solving for the steady state, the linear system with a matrix given as a Kronecker sum is the operation with higher computational cost. A similar bottleneck is encountered for the treatment of the time-dependent problem by implicit methods, which are unavoidable due to the stiffness of the Laplace operator.

The case $\alpha = 1$, which corresponds to the classical Laplace operator, has been analyzed in detail in the literature (see [24] and the references therein). When $d = 2$ the problem can be recast as solving a linear matrix Eq. [30] (called *Lyapunov equation* if $A_1 = A_2$, and *Sylvester equation* otherwise). These equations are often studied by reshaping the vectors x and c into matrices X and C, which yields

$$X A_1^T + A_2 X = C, \qquad c = \mathrm{vec}(C), \quad x = \mathrm{vec}(X).$$

Here, the vec operator stacks all the columns of a matrix on top of the other. In several instances of this problem, the right-hand side matrix C is low-rank, or at least numerically low-rank (i.e., with decaying singular values). This is the case when the right-hand side is the discretization of a (piece-wise) smooth function [35]. Under this assumption, the low-rank property is numerically inherited by the solution X, which can be efficiently approximated using low-rank solvers for matrix equations such as rational Krylov methods [30, 33] or ADI [6].

When $d > 2$, similar results can be obtained, but the derivation is more challenging. In this context one can naturally reshape the vectors x and c as d-dimensional tensors, for which several (non-equivalent) definitions of rank are available [15]. Low-rank approximability results for tensors are given in [24], relying on exponential sum approximations.

Krylov projection methods can be extended to the case $0 < \alpha < 1$ when $d = 2$, using the formulation of the problem as the evaluation of a bivariate matrix function [22, 23, 26, 27]. Although in principle this approach may be used for higher d as well, it leads to multivariate matrix functions and Tucker tensor approximation, which has

an exponential storage and complexity cost in d, and hence does not solve the so-called "curse of dimensionality" [29].

Extending results for tensor Sylvester equations to the case $\alpha < 1$ is inherently difficult since the separability of the operator is lost, and all strategies based on displacement ranks [5, 8, 32] are not easily applicable.

In this work, we consider the use of exponential sums to derive low-rank approximability results and low-rank solvers for the case of a generic d and $0 < \alpha < 1$. Our results can be interpreted as an extension of the exponential sum approximation for $1/z$, see for instance [16] and the reference therein.

The work is structured as follows. In Sect. 2 we derive an exponential sum approximation for $z^{-\alpha}$ over $[1, +\infty]$, and provide guaranteed and explicit error bounds. We prove that this can be used to approximate the solution of the linear systems $\mathcal{A}^{\alpha} x = c$ in a cheap way. In Sect. 3 we show that this representation of the solution can be used to derive approximation results for the solution in tensors in the same low-rank structure used for the right-hand side (Tucker, Tensor-Train, …). We conclude with some numerical experiments in Sect. 4, and draw some final remarks in Sect. 5.

2 Exponential sums

We consider the approximation problem of determining α_j, β_j such that

$$\xi^{-\alpha} \approx \sum_{j=1}^{k} \alpha_j e^{-\beta_j \xi}. \qquad \xi \in [1, \infty). \tag{2}$$

Finding an expression in the above form (which we call *exponential sum*) allows to approximate the function $z^{-\alpha}$ of a matrix $\mathcal{A}$ expressed as a Kronecker sum at a low computational cost. Indeed, if two matrices A and B commute, we have $e^{AB} = e^{BA} = e^A e^B$ [19, Theorem 10.2]. Since all summands in a Kronecker sum commute we can write

$$e^{-\beta \mathcal{A}} = \bigotimes_{i=1}^{d} e^{-\beta A_i}, \qquad \mathcal{A} = \bigoplus_{i=1}^{d} A_i.$$

As we will see in Sect. 3, this is key in deriving low-rank approximability bounds. We rewrite $\xi^{-\alpha}$ in integral form as follows:

$$\xi^{-\alpha} = \frac{1}{\Gamma(\alpha)} \int_0^{\infty} \frac{e^{-t\xi}}{t^{1-\alpha}}\, dt, \qquad \xi \in \mathbb{R}_+. \tag{3}$$

Employing any quadrature rule for approximating (3) yields an approximant of $\xi^{-\alpha}$ by taking a weighted average of evaluations of the integrand, which is exactly in the form of Eq. (2). Let w_j and t_j, for $j = 1, \ldots, k$, be the weights and nodes of such quadrature rule, respectively. Then,

$$\xi^{-\alpha} \approx \sum_{j=1}^{k} w_j \frac{e^{-t_j \xi}}{t_j^{1-\alpha}} = \sum_{j=1}^{k} \alpha_j e^{-\beta_j \xi}, \qquad \begin{cases} \alpha_j = w_j \frac{t_j^{\alpha-1}}{\Gamma(\alpha)} \\ \beta_j = t_j \end{cases}.$$

Our aim is to derive a quadrature that is uniformly accurate over $[1, +\infty)$. We will achieve this goal by a technique called *sinc quadrature*, also known as infinite trapezoidal rule, coupled with appropriate change of variables.

We briefly recap the classical results on sinc quadrature in Sect. 2.1; then, we build the approximation over $[1, \infty)$ in Sect. 2.2, and we show how this can be used to approximate the solution of the linear system $\mathcal{A}^{\alpha} x = c$ and to provide theoretical predictions of approximability in low-rank tensor formats for x, under the assumption that c is itself of low tensor rank (up to appropriately reshaping it).

2.1 Sinc quadrature

We refer the reader to [34] for a more detailed description of the results on sinc quadrature, and in particular [16, Appendix D] for a similar derivation applied to $g(z) := z^{-1}$.

Let $d > 0$ and $g(z)$ be analytic over the infinite strip $\mathcal{D}_d := \{z \mid -d < \Im(z) < d\}$, and such that the integral on the boundary of $\mathcal{D}_d$ is finite, i.e.,

$$\|g\|_{\mathcal{D}_d} := \int_{\partial \mathcal{D}_d} |g(z)| \cdot |dz| < \infty. \tag{4}$$

A sinc quadrature formula is obtained by approximating the integral of $g(z)$ over the real axis by an infinite trapezoidal rule with step h:

$$\int_{-\infty}^{\infty} g(t)\, dt \approx h \sum_{j \in \mathbb{Z}} g(jh).$$

For $h \to 0$, this quadrature converges exponentially. The constant in front of the convergence bound depends on the integral in (4). More precisely, we have

Theorem 1 ([10, Theorem D.26]) *Let $g(z)$ be holomorphic over $\mathcal{D}_d$. Then,*

$$\left| \int_{-\infty}^{\infty} g(t)\, dt - h \sum_{j \in \mathbb{Z}} g(jh) \right| \leq \|g\|_{\mathcal{D}_d} \cdot e^{-2\pi d/h}.$$

The above result is not of immediate practical use, since the discretization of the integral requires to evaluate an infinite series. However, if $g(t)$ decays quickly enough for $|t| \to \infty$, we can truncate the sum and estimate the error by bounding the magnitude of the dropped terms.

To obtain an efficient evaluation scheme, we need to balance the error performed when truncating the series with the one coming from the quadrature rule. Hence, the choice of the number of terms to consider automatically implies an optimal step size h in most cases. This will be discussed in further detail in the next sections.

2.2 Approximating $z^{-\alpha}$ over $[1, \infty)$

The integral form of $\xi^{-\alpha}$ that we considered in Eq. (3) is defined by an integral over $[0, \infty)$. This is not suitable for employing sinc quadrature techniques, and therefore

we need to remap it as an integral over $\mathbb{R}$. To this aim, we introduce the change of variable $t = \log(1+e^{\tau})^{\frac{1}{\alpha}}$; by direct computation, we obtain:

$$\xi^{-\alpha} = \frac{1}{\Gamma(\alpha)} \int_0^{\infty} \frac{e^{-t\xi}}{t^{1-\alpha}}\, dt = \frac{1}{\alpha\Gamma(\alpha)} \int_{-\infty}^{\infty} \frac{e^{-\log(1+e^{\tau})^{\frac{1}{\alpha}}\xi}}{1+e^{-\tau}}\, d\tau. \tag{5}$$

For the sake of notational simplicity, we now define the following shorthand for the integrand:

$$g(\tau) := \frac{e^{-\log(1+e^{\tau})^{\frac{1}{\alpha}}\xi}}{1+e^{-\tau}}$$

We note that $g(\tau)$ implicitly depends on ξ, but we do not report this dependency explicitly to keep the notation more readable. Recall that $\mathcal{D}_d := \{z \mid |\Im(z)| \leq d\} \subseteq \mathbb{C}$ denotes the infinite horizontal strip of width $2d$, centered around the real line.

To use the results on sinc approximation, we first need to ensure that the integrand is analytic on the infinite strip $\mathcal{D}_d$, for suitable choices of d.

Lemma 1 *The function $g(\tau)$ is analytic on $\mathcal{D}_d$ for any $d < \pi$.*

Proof To ensure the analyticity of the integrand $g(\tau)$ we choose to exclude points where $e^{-\tau} = -1$, which force the denominator to vanish, to exclude points $1 + e^{\tau} \in \mathbb{R}_-$, which force the logarithm to be evaluated at its branch cut, and to exclude all points in $\mathbb{R}^-$ from the argument of the fractional power, to avoid the analogous problem for the logarithm implicitly defining it. If these three conditions are met, then the function is obtained through compositions of functions that are analytic on the domain of interest.

We shall deal with these cases separately. The first condition is linked with a class of poles encountered for $\tau = i(2k+1)\pi$, for any $k \in \mathbb{Z}$, and we can exclude them by requiring $d < \pi$. Similarly, this condition automatically implies that $1 + e^{\tau} \notin \mathbb{R}_-$, which excludes evaluations of $\log(1+e^{\tau})$ on its branch cut.

The third situation is encountered when $\log(1+e^{\tau}) \in (-\infty, 0]$, which in turn implies $e^{\tau} \in [-1, 0)$. If we write $\tau = \alpha + i\beta$, this only happens when

$$\alpha \leq 0, \qquad \beta = (2k+1)\pi,\ k \in \mathbb{Z}.$$

As in the previous case, we can avoid this situation by imposing a constraint on d, and requiring $|\beta| \leq d < \pi$. □

We now derive a bound for the integral of the modulus of $g(\tau)$ in (5) over $\partial\mathcal{D}_d$. This imposes further constraints on the choice of d, which are stronger than the ones imposed by Lemma 1. We make the following claims, which will be detailed in this section. Let $z = \gamma \pm id$ be a point in $\partial\mathcal{D}_d$ and $0 \leq d < \frac{\alpha\pi}{4}$. Then,

$$|g(\gamma \pm id)| \leq \begin{cases} e^{-|\gamma|} & \gamma \leq 0 \\ e^{-\xi|\gamma|^{\frac{1}{\alpha}} \cos\left(\frac{d}{\alpha \max\{\gamma, \frac{1}{2}\}}\right)} & \gamma \geq 0 \end{cases}.$$

We prove these results in Lemma 2 and 3, that will be later leveraged to prove the convergence of the exponential sum approximation. We shall see that in order to combine the hypotheses of these results, we will need to ensure that d is chosen as $d \leq \alpha\pi/8$.

Lemma 2 *Let $\tau = \gamma \pm id$, and let $\gamma \leq 0$, $0 < \alpha < 1$, $\xi > 0$ and $0 \leq d \leq \frac{\pi}{2}$ be real numbers such that*

$$\sin d \leq \frac{1}{4} \tan\left(\frac{\alpha\pi}{2}\right). \tag{6}$$

Then,

$$|g(\tau)| \leq \left| \frac{1}{1+e^{-\tau}} \right| \leq e^{-|\gamma|}.$$

Proof To prove the result we show that

$$\left| e^{-\xi \log(1+e^{\gamma \pm id})^{\frac{1}{\alpha}}} \right| \leq 1. \tag{7}$$

If the above condition is satisfied, using $d \leq \frac{\pi}{2}$ we have

$$|g(\tau)| \leq \frac{1}{|1+e^{-\gamma \pm id}|} = \frac{1}{\sqrt{1+e^{-2\gamma}+2e^{-\gamma}\cos d}} \leq e^{-|\gamma|}.$$

We now prove the claim in Eq. (7). Using polar coordinates we can write

$$\log(1+e^{\gamma \pm id}) = \sqrt{\frac{1}{4}\log(1+e^{2\gamma}+2e^{\gamma}\cos d)^2 + \arctan\left(\frac{e^{\gamma}\sin d}{1+e^{\gamma}\cos d}\right)^2} \cdot e^{i\theta(\gamma)},$$

where

$$\theta(\gamma) := \arctan\left(\frac{\pm 2\arctan\left(\frac{e^{\gamma}\sin d}{1+e^{\gamma}\cos d}\right)}{\log(1+e^{2\gamma}+2e^{\gamma}\cos d)}\right).$$

We can write the $\frac{1}{\alpha}$th power of the logarithm as

$$\log(1+e^{\gamma \pm id})^{\frac{1}{\alpha}} = \left(\frac{1}{4}\log(1+e^{2\gamma}+2e^{\gamma}\cos d)^2 + \arctan\left(\frac{e^{\gamma}\sin d}{1+e^{\gamma}\cos d}\right)^2\right)^{\frac{1}{2\alpha}} \cdot e^{i\frac{\theta(\gamma)}{\alpha}}.$$

Since $\xi > 0$, it is sufficient to prove that the real part of the above expression is positive. This is equivalent to imposing that $\cos\left(\frac{\theta(\gamma)}{\alpha}\right) \geq 0$. In particular, we show

that

$$\left|\frac{\theta(\gamma)}{\alpha}\right| = \frac{1}{\alpha}\arctan\left(\frac{2\arctan\left(\frac{e^{\gamma}\sin d}{1+e^{\gamma}\cos d}\right)}{\log(1+e^{2\gamma}+2e^{\gamma}\cos d)}\right) \leq \frac{\pi}{2}. \tag{8}$$

The second inequality is equivalent to imposing

$$\frac{2\arctan\left(\frac{e^{\gamma}\sin d}{1+e^{\gamma}\cos d}\right)}{\log(1+e^{2\gamma}+2e^{\gamma}\cos d)} \leq \tan\left(\frac{\pi\alpha}{2}\right).$$

Recalling that $\arctan(x) \leq x$ for all $x \geq 0$, we have

$$\arctan\left(\frac{e^{\gamma}\sin d}{1+e^{\gamma}\cos d}\right) \leq \frac{e^{\gamma}\sin d}{1+e^{\gamma}\cos d} \leq e^{\gamma}\sin d$$

and using the inequality $\log(1+x) \geq x - \frac{1}{2}x^2$ for $x \geq 0$, we have

$$\log(1+e^{2\gamma}+2e^{\gamma}\cos d) \geq \log(1+e^{\gamma}) \geq e^{\gamma} - \frac{1}{2}e^{2\gamma}.$$

Hence,

$$\frac{2\arctan\left(\frac{e^{\gamma}\sin d}{1+e^{\gamma}\cos d}\right)}{\log(1+e^{2\gamma}+2e^{\gamma}\cos d)} \leq \frac{2\sin d}{1-\frac{1}{2}e^{\gamma}} \leq 4\sin d.$$

We conclude by using the hypothesis (6), which implies that the right-hand side is bounded by $\tan\left(\frac{\pi\alpha}{2}\right)$, as needed. □

The next result controls the magnitude of the integrand when the real part of the integration variable $\tau = \gamma \pm id$ is positive, which enables to bound the norm of the integral in the right half plane.

Lemma 3 *Let $\tau = \gamma \pm id$, and let $\gamma > 0, \xi > 0$, $0 < \alpha < 1$, and $0 \leq d < \frac{\alpha\pi}{4}$ be real numbers; then, the function $g(\tau)$ is bounded above in modulus by*

$$|g(\tau)| \leq \left|e^{-\xi\log(1+e^{\gamma\pm id})^{\frac{1}{\alpha}}}\right| \leq e^{-\xi|\gamma|^{\frac{1}{\alpha}}\cos\left(\frac{d}{\alpha\max\{\gamma,\frac{1}{2}\}}\right)}.$$

Proof We may write

$$|g(\tau)| = \frac{\left|e^{-\xi\log(1+e^{\gamma\pm id})^{\frac{1}{\alpha}}}\right|}{|1+e^{-\tau}|} \leq \frac{\left|e^{-\xi\log(1+e^{\gamma\pm id})^{\frac{1}{\alpha}}}\right|}{|1+e^{-\gamma}\cos d|} \leq \left|e^{-\xi\log(1+e^{\gamma\pm id})^{\frac{1}{\alpha}}}\right|,$$

thanks to $\cos d \geq 0$. We now prove the second inequality; thanks to $\xi \in \mathbb{R}$,

$$\left| e^{-\xi \log(1+e^{\gamma \pm id})^{\frac{1}{\alpha}}} \right| = \exp(-\xi \Re(\log(1+e^{\gamma \pm id})^{\frac{1}{\alpha}})).$$

Hence, in order to devise an upper bound for the left-hand side, we need a lower bound for the real part of the logarithm in the right-hand side. By writing the argument of the logarithm in polar coordinates we obtain the following expression:

$$\log(1+e^{\gamma \pm id}) = \frac{1}{2}\log(1+e^{2\gamma}+2e^{\gamma}\cos d) \pm i \arctan\left(\frac{e^{\gamma}\sin d}{1+e^{\gamma}\cos d}\right).$$

We now rewrite the above in polar coordinates, which yields

$$\log(1+e^{\gamma \pm id}) = \sqrt{\frac{1}{4}\log(1+e^{2\gamma}+2e^{\gamma}\cos d)^2 + \arctan\left(\frac{e^{\gamma}\sin d}{1+e^{\gamma}\cos d}\right)^2} \cdot e^{i\theta(\gamma)},$$

where

$$\theta(\gamma) := \arctan\left(\frac{\pm 2 \arctan\left(\frac{e^{\gamma}\sin d}{1+e^{\gamma}\cos d}\right)}{\log(1+e^{2\gamma}+2e^{\gamma}\cos d)}\right).$$

This gives an explicit formula for the real part of the above logarithm raised to the power $\frac{1}{\alpha}$:

$$\Re(\log(1+e^{\gamma \pm id})^{\frac{1}{\alpha}}) = \left[\frac{1}{4}\log(1+e^{2\gamma}+2e^{\gamma}\cos d)^2 + \arctan\left(\frac{e^{\gamma}\sin d}{1+e^{\gamma}\cos d}\right)^2\right]^{\frac{1}{2\alpha}} \cos\left(\frac{\theta(\gamma)}{\alpha}\right).$$

The above yields an exact expression for the quantity that we need to bound. We now make some simplifications, employing the following inequalities:

$$\log(1+e^{2\gamma}+2e^{\gamma}\cos d) \geq \max\{2\gamma, 1\} \quad 0 \leq \arctan\left(\frac{e^{\gamma}\sin d}{1+e^{\gamma}\cos d}\right) \leq d. \tag{9}$$

The two inequalities can be combined to show that

$$0 \leq \theta(\gamma) \leq \frac{d}{\max\{\gamma, \frac{1}{2}\}} \implies \cos\left(\frac{\theta(\gamma)}{\alpha}\right) \geq \cos\left(\frac{d}{\alpha \max\{\gamma, \frac{1}{2}\}}\right),$$

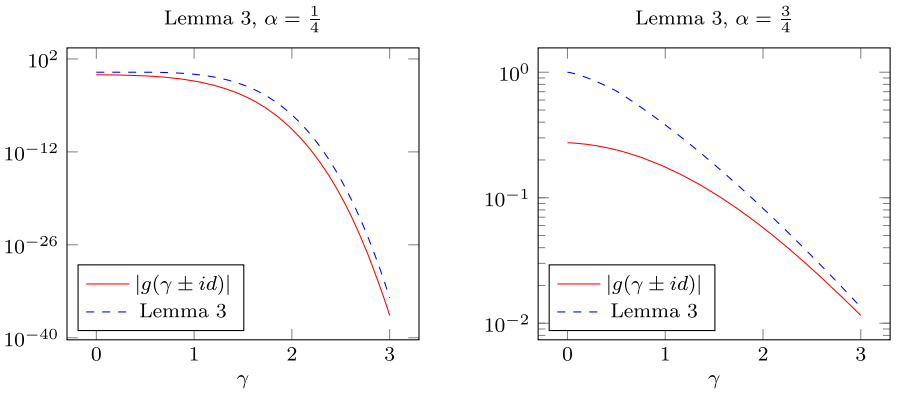

Fig. 1 Bounds for the modulus of $g(\tau)$, obtained for $\alpha \in \{\frac{1}{4}, \frac{3}{4}\}$ by Lemma 3. The value of d in these examples is chosen as $d = \frac{\pi}{16}$

where we used that $0 \leq \theta(\gamma) \leq \frac{\alpha\pi}{2}$ in view of $d \leq \frac{\pi\alpha}{4}$. We now make use again of (9) to bound the first factor, obtaining

$$\Re(\log(1 + e^{\gamma \pm id})^{\frac{1}{\alpha}}) \geq \gamma^{\frac{1}{\alpha}} \cos\left(\frac{d}{\alpha \max\{\gamma, \frac{1}{2}\}}\right),$$

which implies the sought bound. □

Even though we have made some simplifications in the expressions, the asymptotic behavior for $\gamma \to \pm\infty$ is tight. In addition, for the values of γ close to zero, the bound is still quite descriptive of the actual behavior, as we show in Fig. 1 for a few values of α.

We now have all the tools to give an explicit upper bound for the integral of the modulus of $g(\tau)$ over the boundary of $\mathcal{D}_d$.

Lemma 4 *For any d satisfying $0 < d \leq \frac{\pi\alpha}{8}$ with $0 < \alpha < 1$, it holds:*

$$\|g\|_{\mathcal{D}_d} = \int_{\partial\mathcal{D}_d} |g(\tau)| \cdot |d\tau| \leq 2\left(1 + \log(2) + \frac{\Gamma(\alpha + 1)}{(\xi \cos(\frac{\pi}{8}))^{\alpha}}\right).$$

Proof First, we note that for any d in the region of interest we have

$$\sin d \leq d \leq \frac{\pi\alpha}{8} \leq \frac{1}{4}\tan\left(\frac{\pi\alpha}{2}\right),$$

and therefore the hypotheses of Lemma 2 and 3 are satisfied. In addition, thanks to the property $|g(\tau)| = |g(\overline{\tau})|$, we may rewrite the integral as

$$\int_{\partial\mathcal{D}_d} |g(\tau)| \cdot |d\tau| = 2\int_0^{\infty} |g(\gamma + id)| d\gamma + 2\int_{-\infty}^{0} |g(\gamma + id)| d\gamma.$$

The integrands can be dealt with separately. In $(-\infty, 0]$ we can use Lemma 2 to obtain the following bound:

$$2\int_{-\infty}^{0} |g(\gamma + id)|d\gamma \le 2\int_{-\infty}^{0} \frac{1}{1+e^{-\tau}} d\gamma = 2\log(2).$$

Similarly, we can bound the integral from 0 to ∞ using Lemma 3 as follows:

$$\begin{aligned} 2\int_{0}^{\infty} |g(\gamma + id)|d\gamma &\le 2\int_{0}^{1} d\gamma + 2\int_{1}^{\infty} e^{-\xi|\gamma|^{\frac{1}{\alpha}}\cos(\frac{d}{\alpha})} d\gamma \\ &\le 2 + \frac{2\Gamma(\alpha+1)}{(\xi\cos(\frac{d}{\alpha}))^{\alpha}} =\le 2 + \frac{2\Gamma(\alpha+1)}{(\xi\cos(\frac{\pi}{8}))^{\alpha}}, \end{aligned}$$

where in the last inequality we have used $d \le \frac{\pi\alpha}{8}$. The result follows by combining these two bounds. □

Remark 1 The bound for the integrand in $[0, \infty)$ is not asymptotically sharp, since for $\gamma \to \infty$ we have $\cos(\frac{d}{\alpha\gamma}) \to 1$, and instead we have replaced it with $\cos(\pi/8) \approx 0.9238795\ldots$ in the proof of Lemma 4; however, this does not make a dramatic difference in practice, and makes the result much more readable.

Thanks to the estimate of Lemma 4, we may now approximate $\xi^{-\alpha}$ with an infinite series as follows:

$$\xi^{-\alpha} = h\sum_{j\in\mathbb{Z}} g(jh) + \epsilon_h, \qquad |\epsilon_h| \le 2\left(1 + \log(2) + \frac{\Gamma(\alpha+1)}{(\xi\cos(\frac{\pi}{8}))^{\alpha}}\right) e^{-2\pi d/h}.$$

However, this does not yet give us a practical algorithm, since we need to truncate the series to a finite sum. We use the following notation:

$$E(g,h) = h\sum_{j\in\mathbb{Z}} g(jh) \quad \text{and} \quad E_{N_-,N_+}(g,h) = h\sum_{j=-N_-}^{N_+} g(jh).$$

We need an estimate for the error introduced by truncating the sum to N_+ positive terms and N_- negative ones. We state the following lemma, which is tailored to the decay properties of the function $g(\tau)$ considered in this section.

Lemma 5 *Let c_-, c_+ and β be positive constants such that*

$$g(x) \le c_- e^{-|x|} \quad \textit{for } x \le 0, \tag{10}$$

and

$$g(x) \le c_+ e^{-\beta|x|^{\frac{1}{\alpha}}} \quad \textit{for } x \ge 0. \tag{11}$$

Then, the remainder $E(g,h) - E_{N_-,N_+}(g,h)$ *satisfies:*

$$\left|E(g,h) - E_{N_-,N_+}(g,h)\right| \le c_- \frac{e^{-N_- h}}{h} + c_+ \frac{\alpha e^{-\beta(N_+ h)^{\frac{1}{\alpha}}}}{\beta h^{\frac{1}{\alpha}}}. \tag{12}$$

Proof Since

$$E(g,h) - E_{N_-,N_+}(g,h) = \sum_{k>N_-} g(-kh) + \sum_{k>N_+} g(kh),$$

using (10) and (11) we have

$$\begin{aligned} E(g,h) - E_{N_-,N_+}(g,h) &\le c_- \sum_{k>N_-} e^{-kh} + c_+ \sum_{k>N_+} e^{-\beta(kh)^{\frac{1}{\alpha}}} \\ &\le c_- \int_{N_-}^{\infty} e^{-kh} dk + c_+ \int_{N_+}^{\infty} e^{-\beta(kh)^{\frac{1}{\alpha}}} dk \\ &= c_- \frac{e^{-N_- h}}{h} + c_+ \int_{k>N_+}^{\infty} e^{-\beta(kh)^{\frac{1}{\alpha}}} dk. \end{aligned}$$

To give an upper bound to the last integral let $x = k^{\frac{1}{\alpha}}$. We then have

$$\int_{k>N_+}^{\infty} e^{-\beta(kh)^{\frac{1}{\alpha}}} dk = \alpha \int_{x>N_+^{\frac{1}{\alpha}}}^{\infty} \frac{e^{-\beta x h^{\frac{1}{\alpha}}}}{x^{1-\alpha}} \le \alpha \int_{x>N_+^{\frac{1}{\alpha}}}^{\infty} e^{-\beta x h^{\frac{1}{\alpha}}} = \alpha \frac{e^{-\beta(N_+ h)^{\frac{1}{\alpha}}}}{\beta h^{\frac{1}{\alpha}}}.$$

□

We now address the problem of determining the number of terms required to have a prescribed accuracy ϵ. Theorem 1 suggests that h should be chosen to have $e^{-2\pi d/h} \approx \epsilon$. If N_+ and N_- are chosen as requested by the next theorem, this guarantees the required accuracy.

Theorem 2 *Let* $\epsilon > 0$ *and* $0 < \alpha < 1$. *Then, for any* $0 < d \le \frac{\pi\alpha}{8}$, *and*

$$h = \frac{2\pi d}{\log(\epsilon^{-1})}, \qquad N_- = \frac{2\pi d}{h^2}, \qquad N_+ = \left(\frac{2\pi d h^{-\frac{\alpha+1}{\alpha}}}{\beta}\right)^{\alpha},$$

where $\beta = \cos(2d/\alpha) \ge \cos(\pi/4)$, *it holds*

$$|\xi^{-\alpha} - E_{N_-,N_+}(g,h)| \le \left(\|g\|_{\mathcal{D}_d} + \frac{1}{h} + \frac{1}{\beta h^{\frac{1}{\alpha}}}\right)\epsilon.$$

If ϵ is chosen smaller than $e^{-\pi^2/4} \approx 0.085$, then the error can be bounded by

$$|\xi^{-\alpha} - E_{N_-,N_+}(g,h)| \leq 2\left[1 + \log(2) + \frac{\Gamma(\alpha+1)}{\cos(\pi/8)^\alpha} + \cos(\pi/4)^{-1}\left(\frac{4\log(\epsilon^{-1})}{\pi^2\alpha}\right)^{\frac{1}{\alpha}}\right]\epsilon.$$

In particular, for large N and small ϵ, the asymptotics $\epsilon \sim \mathcal{O}(e^{-\sqrt{2\pi d N}})$ and $N \sim \mathcal{O}(\log^2(\frac{1}{\epsilon})/2\pi d)$ hold up to logarithmic factors.

Proof Leveraging Theorem 1, we can bound the quadrature error by

$$|\xi^{-\alpha} - E(g,h)| \leq \|g\|_{\mathcal{D}_d} e^{-2\pi d/h}.$$

We now show that the proposed choices of N_- and N_+ provide an error bound with the same exponential convergence, but different constants in front. Using Lemma 5 we obtain:

$$\begin{aligned}|E(g,h) - E_{N_-,N_+}(g,h)| &\leq c_- \frac{e^{-N_- h}}{h} + c_+ \frac{\alpha e^{-\beta(N_+ h)^{\frac{1}{\alpha}}}}{\beta h^{\frac{1}{\alpha}}}. \\ &\leq \left(\frac{c_-}{h} + \frac{\alpha c_+}{\beta h^{\frac{1}{\alpha}}} +\right) e^{-2\pi d/h},\end{aligned}$$

where $\beta = \cos(2d/\alpha)$ applying Lemma 3 with the inequality $\max\{\frac{1}{2}, \gamma\} \geq \frac{1}{2}$. We have that $c_- = 1$ thanks to Lemma 2 and $c_+ = 1$, thanks to Lemma 3. The final bound is obtained using the explicit expression for $\|g\|_{\mathcal{D}_d}$ together with

$$\epsilon \leq e^{-\pi^2/4} \implies h = \frac{2\pi d}{\log(\epsilon^{-1})} \leq \frac{\pi^2\alpha}{4\log(\epsilon^{-1})} \leq \frac{\pi^2}{4\log(\epsilon^{-1})} \leq 1,$$

which implies $\frac{1}{\beta h^{\frac{1}{\alpha}}} \geq \frac{1}{h}$, and therefore allows to give the upper bound

$$\frac{1}{h} + \frac{1}{\beta h^{\frac{1}{\alpha}}} \leq \frac{2}{\beta h^{\frac{1}{\alpha}}} \leq \frac{2}{\cos(\pi/4)}\left(4\frac{\log(\epsilon^{-1})}{\pi^2\alpha}\right)^{\frac{1}{\alpha}}.$$

The claim on the asymptotic growth for $\epsilon, h \to 0$ follows by noting that the dominant term in N is $N_- \sim \mathcal{O}(h^{-2})$. □

We now verify the convergence predicted by these results by considering different $\xi \in [1, \infty)$, logarithmically spaced on $[1, 10^6]$. For these values, we compute the exponential sum approximating $\xi^{-\alpha}$ for $\alpha \in \{0.25, 0.75\}$. The results, including the asymptotic bound from Theorem 2, are reported in Fig. 2. It is visible how Theorem 2 accurately describes the asymptotic rate of convergence of the approximation.

We note that to reach machine precision, a non trivial amount of exponentials has to be computed. In addition, when α is small d has to be chosen small as well, obtaining a slower convergence speed, as predicted by Theorem 2.

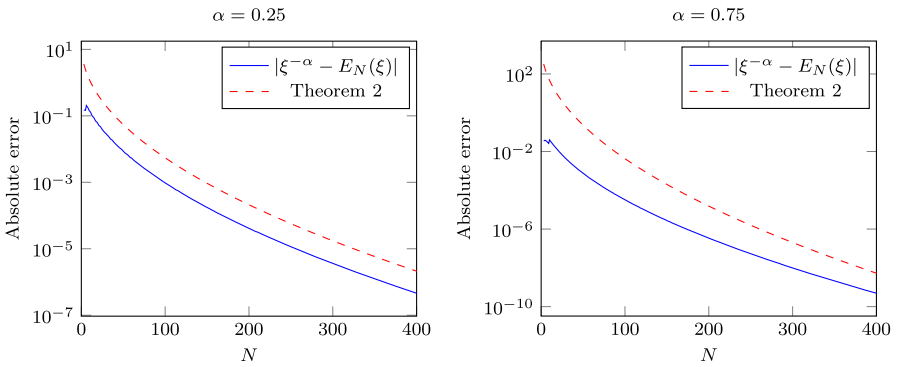

Fig. 2 Absolute errors of the exponential sum approximation for ξ^{α}. The error is the maximum for 100 logarithmically spaced samples over $[1, 10^6]$. The value of α is chosen as 0.25 and 0.75. The convergence speed for $0.25 \leq \alpha \leq 0.75$ interpolates these two examples

3 Low-rank approximability

We now make use of the results developed in Sect. 2 to prove that the solutions of Kronecker-structured linear systems inherit the low-rank tensor structure of the right-hand side. Recall that we are interested in linear systems of the form:

$$\mathcal{A}^{\alpha} x = c, \qquad \mathcal{A} = \bigoplus_{i=1}^{d} A_i, \qquad A_i \in \mathbb{C}^{n_i \times n_i}. \tag{13}$$

where as in (1) the "$\oplus$" symbol denotes the Kronecker sum.

The vectors x and c may be naturally reshaped into $n_1 \times \ldots \times n_d$ tensors; we denote these reshaped versions with the capital letters X, C, respectively; we will use this notation throughout the section; for instance, for the vector x (and tensor X), we have the correspondence:

$$X \in \mathbb{C}^{n_1 \times \ldots \times n_d} \longleftrightarrow x = \mathrm{vec}(X) \in \mathbb{C}^{n_1 \cdots n_d}.$$

The linear system (13) can be rephrased as computing $x = f(\mathcal{A})c$, where $f(z) = z^{-\alpha}$, and has therefore a very natural connection with the exponential sum approximation that we have discussed in the previous section.

When dealing with high-dimensional problems (i.e., the integer d is large) it is natural to assume that some low-rank structure is present in the tensor C. If this assumption is not satisfied, it is unlikely that storing C is possible at all.

For analogous reasons, we need to guarantee that X is endowed with a similar structure: otherwise, there is little hope of computing it, if there is not sufficient storage for memorizing it. The exponential sum approximation can be used to guarantee that X inherits the low-rank structure from the right-hand side C and this is precisely the goal of this section.

In contrast to what happens with matrices, there are many competing definitions of a low-rank tensor. In this work, we consider tensors with low CP-rank, TT-rank, and multilinear rank [15].

We briefly recall the definition and properties of these families in Sect. 3.1, and then show the results obtainable through the exponential sum approximation in Sect. 3.

3.1 Low-rank tensor formats

A natural way to define the rank of a d-dimensional tensor $X \in \mathbb{C}^{n_1 \times \dots \times n_d}$ is as the *minimum length of a "low-rank decomposition"*, here written for simplicity on the vectorization $x = \mathrm{vec}(X)$:

$$x = u_{1,1} \otimes \cdots \otimes u_{1,d} + \cdots + u_{k,1} \otimes \cdots \otimes u_{k,d}.$$

This is usually called just *tensor rank* or *CP rank*, and the above decomposition is called a *Canonical Polyadic Decomposition* (CPD or CP decomposition). Despite its simplicity, computing such decomposition is numerically challenging for large d [21], in contrast to what happens when $d = 2$, when we can leverage the singular value decomposition (SVD).

For this reason, several alternative definitions of low-rank tensors (and the associated decompositions) have been introduced in recent years. We mention the multilinear singular value decomposition [12], often shortened as HOSVD (High Order SVD), and the tensor train format [28]. Both these formats have an SVD-like procedure that allows to obtain the best (or at least quasi-optimal) low-rank approximation to a tensor X. To discuss the properties of these formats, we shall introduce the definition of unfolding.

Definition 1 The ith mode unfolding $X^{(i)}$ of a tensor X is the matrix obtained by stacking the vectors $X^{(i)}_j$ containing the entries of X with the ith index equal to j for $j = 1, \dots, n_i$, i.e.,

$$X^{(i)} = \begin{bmatrix} \mathrm{vec}(X^{(i)}_1)^T \\ \vdots \\ \mathrm{vec}(X^{(i)}_{n_i})^T \end{bmatrix} \in \mathbb{C}^{n_i \times (n/n_i)},$$

where $n = \prod_{i=1}^{d} n_i$.

The unfoldings can be used to define the multilinear rank of a tensor X.

Definition 2 ([12]) The *multilinear rank* of a tensor X is the tuple $r = (r_1, \dots, r_d)$, where $r_i = \mathrm{rank}(X^{(i)})$, and $X^{(i)}$ is the ith mode unfolding of X.

We often say that a tensor has multilinear rank smaller than $r = (r_1, \dots, r_d)$, to mean that the rank is component-wise smaller. We can use matrices to act on tensors, as described by the following.

Definition 3 Given a matrix $A \in \mathbb{C}^{m_j \times n_j}$ and a tensor $X \in \mathbb{C}^{n_1 \times \dots \times n_d}$, the j*th mode product of* X *times* A, denoted by $X \times_j A$, is the d-dimensional tensor $Y \in \mathbb{C}^{n_1 \times \cdots \times n_{j-1} \times m_j \times n_{j+1} \times \cdots \times n_d}$, defined by:

$$Y_{i_1,\dots,i_d} = \sum_{k=1}^{n_j} A_{i_j k} X_{i_1,\dots,i_{j-1},k,i_{j+1},\dots,i_d}.$$

If $d = 2$ and therefore X is a matrix, we have $X \times_1 A = AX$ and $X \times_2 A = XA^T$. Hence, this operation can be seen as the high-order generalization of left and right matrix multiplication. We remark a few useful properties that relate unfoldings and jth mode products.

Lemma 6 *Let $Y = X \times_i A$. Then,*

(i) $Y^{(i)} = AX^{(i)}$;
(ii) $y = (\underbrace{I \otimes \dots \otimes I}_{i-1 \text{ terms}} \otimes A \otimes \underbrace{I \otimes \dots \otimes I}_{d-i-1 \text{ terms}})x$;
(iii) the multilinear rank of Y is bounded by $r = (r_1, \dots, r_d)$, the multilinear rank of X;
(iv) for any other tensor Z with multilinear rank $(s_1, \dots, s_d)$, the multilinear rank of $X + Z$ is bounded by $(r_1 + s_1, \dots, r_d + s_d)$.

where as usual $x = \mathrm{vec}(X)$, $y = \mathrm{vec}(Y)$, and the Kronecker product in (ii) has the only matrix different from the identity in position i.

A direct consequence of the second representation of the ith mode product is that, for any choice of matrices A, B and $i \neq j$, we have $(X \times_j B) \times_i A = (X \times_i A) \times_j B$. Hence, we avoid unnecessary brackets when combining several j-mode products writing $X \times_{j_1} A_{j_1} \dots \times_{j_\ell} A_{j_\ell}$.

The (quasi)-optimal multilinear rank $r = (r_1, \dots, r_d)$ approximant to a generic tensor X can be effectively computed by repeatedly truncating the ith mode unfoldings; this procedure is usually known as *multilinear SVD* [12].

If a tensor X has a low multilinear rank, it can be efficiently expressed through a *Tucker decomposition*; with our current notation this can be written as follows:

$$X = B \times_1 U_1 \times_2 U_2 \dots \times_d U_d,$$

where $B \in \mathbb{C}^{r_1 \times \dots \times r_d}$, and U_j are $n_j \times r_j$ matrices with orthogonal columns. When the multilinear ranks are smaller than the dimensions $n_1, \dots, n_d$, this representation allows to compress the data.

We remark that for very large d, this representation can still be too expensive: even if the r_i are small, the storage requirement depends on their product; making the simplifying assumption that $r := r_1 = \dots = r_d$ the storage requirements for this decomposition are $\mathcal{O}(r^d + (n_1 + \dots + n_d)r)$ memory — which is exponential with respect to d. So, even if when $r \ll n_i$ this format allows to save a large amount of memory, working with general high-dimensional problems may remain unfeasible.

To overcome this drawback, several other tensor formats have been introduced: *Tensor Trains* [28] (also called *Matrix Product States*, or MPS [31]), *Hierarchical Tucker Decompositions* [14], and more general *Tensor Networks* [11].

We now briefly recap the properties of Tensor-Trains that are relevant for our results. The TT format requires another definition of rank (the TT-ranks), which requires

the introduction of appropriate matricizations. We expect low-rank approximability properties analogous to the one that we prove for the TT format to hold for other formats as well (such as Hierarchical Tucker, or Tensor Networks [14]).

Given a d-dimensional tensor X, we define the matrices $X^{\{i\}}$ obtained by grouping the first i indices together as row indices, and the remaining ones as columns indices. The vector $r = (r_1, \ldots, r_{d-1})$, where r_i is the rank of $X^{\{i\}}$, is called the *Tensor-Train rank of* X (or TT-rank).

A tensor with TT-rank smaller than $(r_1, \ldots, r_{d-1})$ can be decomposed as follows [28]:

$$X_{i_1,\ldots,i_d} := \sum_{s_1 \le r_1, \ldots, s_{d-1} \le r_{d-1}} C^{(1)}_{i_1 s_1} C^{(2)}_{s_1 i_2 s_2} \ldots C^{(d-1)}_{s_{d-2} i_{d-1} s_{d-1}} C^{(d)}_{s_{d-1} i_d}, \tag{14}$$

where $C^{(j)}$ are called *carriages* and can be either matrices ($j = 1, d$) or three-dimensional tensors ($1 < j < d$). It is readily apparent that this representation breaks the so-called *curse of dimensionality*: a tensor with low (TT-)ranks can be stored with a number of parameters only polynomial in d.

From Eq. (14) we note that the operation $X \times_j A$ can be efficiently evaluated in the TT-format, as that only requires to modify $C^{(j)}$ by computing $C^{(j)} \times_2 A$ (with the only exception $j = 1$, where the required operation is $C^{(1)} \times_1 A$). Hence, we may state a Tensor Train analogue of the last item in Lemma 6.

Lemma 7 *Let $Y = X \times_i A$. Then,*

1. *the Tensor Train rank of Y is bounded by $r = (r_1, \ldots, r_{d-1})$, the Tensor Train rank of X,*
2. *for any other tensor Z with Tensor-Train rank $(s_1, \ldots, s_{d-1})$, the Tensor Train rank of $Y + Z$ is bounded by $(r_1 + s_1, \ldots, r_{d-1} + s_{d-1})$.*

Proof The first claim follows by the current discussion, since the dimensions of the updated carriage $C^{(j)} \times_2 A$ involving the ranks are not modified. We refer the reader to [28, Section 4.1] for a proof of the second one. □

3.2 Low-rank approximation in the symmetric positive definite case

We consider the case where the matrices A_i defining $\mathcal{A}$ are symmetric positive definite. On one hand, this greatly simplifies the derivation of the results thanks to the normality and the fact that the spectrum of $\mathcal{A}$ is real. On the other hand, the non-negativity of the eigenvalues is a common assumption when taking the negative fractional power of an operator, and therefore it is not particularly restrictive in practice.

We will make repeated use of the following fact.

Lemma 8 *Let A_i, $i = 1, \ldots, d$, be matrices of size $n_i \times n_i$, and X any d-dimensional tensor of size $n_1 \times \ldots \times n_d$. Then,*

$$\exp\left(\bigoplus_{i=1}^{d} A_i\right)\mathrm{vec}(X) = \mathrm{vec}(X \times_1 \exp(A_1) \times_2 \ldots \times_d \exp(A_d)).$$

Proof The proof follows noting that the summands defining $\bigoplus_{i=1}^{d} A_i$ commute, and using the property that if $AB = BA$ then $e^{AB} = e^{BA} = e^A e^B$ [19, Theorem 10.2]. □

Theorem 3 *Let $\mathcal{A} = \bigoplus_{i=1}^{d} A_i$ be invertible, with A_i symmetric positive definite matrices. Let $x = \mathrm{vec}(X)$, $c = \mathrm{vec}(C)$, and $x = \mathcal{A}^{-\alpha}c$. Then, for any $N_- \in \mathbb{N}$ satisfying $N_- \geq \frac{\cos(\pi/4)^{1-\alpha}2\alpha}{2\pi d}$ there exists an approximant X_{N_-} to X such that*

$$\|X - X_{N_-}\|_F \leq$$
$$2\lambda_{\min}^{-\alpha}\left[1 + \log(2) + \frac{\Gamma(\alpha+1)}{\cos(\pi/8)^{\alpha}} + \cos(\pi/4)^{-1}\left(\frac{2\sqrt{\alpha N_-}}{\pi\alpha}\right)^{\frac{1}{\alpha}}\right] e^{-\frac{\pi}{2}\sqrt{\alpha N_-}}\|C\|_F.$$

and:

- *if C has CP rank bounded by r, then X_{N_-} has CP rank bounded by $(2N_- + 1)r$.*
- *if C has multilinear rank bounded by $(r_1, \ldots, r_d)$, then the multilinear rank of X_{N_-} is bounded by $((2N_- + 1)r_1, \ldots, (2N_- + 1)r_d)$.*
- *if C has TT-ranks bounded by $(r_1, \ldots, r_{d-1})$ then the approximation X_{N_-} has TT-ranks bounded by $((2N_- + 1)r_1, \ldots, (2N_- + 1)r_{d-1})$.*

Proof Let $f_N(\xi)$ be the exponential sum approximation to $\xi^{-\alpha}$ with $N = N_- + N_+ + 1$ terms of the form

$$\xi^{-\alpha} \approx f_N(\xi) = E_{N_-,N_+}(g,h) = \sum_{j=-N_-}^{N_+} \alpha_j e^{-\beta_j \xi}, \qquad \xi \geq 1.$$

obtained from Theorem 2. Then, we define the approximation X_{N_-} as follows:

$$X_{N_-} = \lambda_{\min}^{-\alpha} \sum_{i=-N_-}^{N_+} \alpha_{i,j} C \times_1 e^{-\beta_{i,j}\lambda_{\min}^{-1}A_1} \ldots \times_d e^{-\beta_{i,j}\lambda_{\min}^{-1}A_d},$$

where $\lambda_{\min}$ is the smallest eigenvalue of $\mathcal{A}$. Let $N = N_- + N_+ + 1$, the amount of terms in the sum. Using the definition of CP rank, and Lemma 6, Lemma 7, we make the following observations:

- If C has CP rank bounded by r, then X_N has CP rank bounded by Nr.
- If the multilinear rank of C is component-wise bounded by $r = (r_1, \ldots, r_d)$, then the multilinear rank of X_N can be controlled with $(Nr_1, \ldots, Nr_d)$ — thanks to Lemma 6.
- If the TT-rank of C is bounded by $r = (r_1, \ldots, r_{d-1})$ then the TT-rank of X_N is bounded by $(Nr_1, \ldots, Nr_{d-1})$ — thanks to Lemma 7.

We now show that the approximation X_{N_-} satisfies the sought bound. First, we show that $N_- \geq N_+$; we have, thanks to the definition of N_- and N_+ in Theorem 2:

$$\frac{N_-}{N_+} = \frac{2\pi d}{h^2}\left[\left(\frac{2\pi d}{\beta}\right)^{\alpha} h^{-(\alpha+1)}\right]^{-1} = (2\pi d)^{1-\alpha} h^{\alpha-1}\beta^{\alpha}.$$

We now have $N_- \geq N_+ \iff N_-/N_+ \geq 1$, which holds if

$$h^{\alpha-1} \geq (2\pi d)^{\alpha-1}\beta^{-\alpha} \iff h \leq (2\pi d)\beta^{\frac{\alpha}{1-\alpha}}.$$

Since $\beta \geq \cos(\pi/4)$, we can instead impose that $h \leq (2\pi d)(\cos(\pi/4))^{\frac{\alpha}{1-\alpha}}$, which is implied by our assumption

$$N_- \geq \frac{(\cos(\pi/4))^{1-\alpha} 2\alpha}{2\pi d}.$$

Therefore, we have that $N \leq 2N_- + 1$. Using the representation $\mathrm{vec}(X) = x = \lambda_{\min}^{-\alpha} f_N(\lambda_{\min}^{-1}\mathcal{A})c$, we obtain

$$\begin{aligned}\|X - X_{N_-}\|_F &= \|\mathcal{A}^{-\alpha}c - \lambda_{\min}^{-\alpha} f_N(\lambda_{\min}^{-1}\mathcal{A})c\|_2 \\ &\leq \lambda_{\min}^{-\alpha}\|(\lambda_{\min}^{-1}\mathcal{A})^{-\alpha} - f_N(\lambda_{\min}^{-1}\mathcal{A})\|_2 \cdot \|c\|_2,\end{aligned}$$

where we have used that $\lambda_{\min}^{-1}\mathcal{A}$ is normal and has spectrum contained in $[1, +\infty)$. We now apply Theorem 2 to obtain

$$\begin{aligned}&\|X - X_{N_-}\|_F \\ &\leq 2\lambda_{\min}^{-\alpha}\left[1 + \log(2) + \frac{\Gamma(\alpha+1)}{(\cos(\pi/8))^{\alpha}} + (\cos(\pi/4))^{-1}\left(\frac{4\sqrt{2\pi d N_-}}{\pi^2\alpha}\right)^{\frac{1}{\alpha}}\right] \\ &\quad e^{-\sqrt{2\pi d N_-}}\|C\|_F.\end{aligned}$$

We can choose $d = \frac{\pi\alpha}{8}$, and obtain the sought result. □

3.3 Connection with rational approximations

In the matrix case ($d = 2$) bounds on the rank of the solution can be obtained by linking the problem with rational approximation on the complex plane. In the special case $\alpha = 1$, this links to the well-known properties of low-rank Sylvester solvers such as ADI, that allows to build (explicit) approximants to the solution X of $AX + XB = C$ in the form

$$X - X_N = r(A)Xr(-B)^{-1}, \qquad r(z) = \frac{p(z)}{q(z)}, \qquad \mathrm{rank}(X_N) \leq N \cdot \mathrm{rank}(C),$$

where $p(z)$ and $q(z)$ are polynomials of degree at most $N + 1$. Considering rational functions which are small on an interval containing the spectrum of A and large on an interval containing the one of B, we can build low-rank approximants to X. The

problem of finding such rational functions is called a *Zolotarev problem*, and the solution (for two real intervals) is known explicitly in terms of elliptic functions [38].

When $\alpha < 1$ the situation is less straightforward because an equation with separable coefficients cannot be written. However, similar results can be derived by using a Cauchy–Stieltjes formulation for $z^{-\alpha}$:

$$z^{-\alpha} = \frac{\sin(\alpha\pi)}{\pi} \int_0^\infty \frac{t^{-\alpha}}{t+z}\, dz.$$

This representation yields a formula for the solution $x = \mathrm{vec}(X)$ to $\mathcal{A}^\alpha x = c$ in terms of integrals of a parameter dependent family of (shifted) Sylvester equations, and this can be used to determine a low-dimensional subspace where a good approximation for the solution can be found. This has been exploited in [7, 26] for constructing rational Krylov methods for the case $d = 2$ and $\alpha < 1$, which predict an exponential decay in the singular values (as opposed to the square root exponential bound from Theorem 2).

Since multilinear and tensor-train ranks are defined by matricization, we think that a similar idea may be exploited to derive bounds for these special cases for $d > 2$, although to the best of our knowledge this has not been worked out explicitly at the time of writing.

A good indication in this direction is given by the numerical experiments, which show a better approximability with respect to these formats than the one predicted by Theorem 3. There is instead little hope to apply such techniques to the CP case.

It is worth mentioning that the connection with rational approximant of $z^{-\alpha}$ have been exploited in many works [1–3, 9, 17, 18] for designing efficient solvers for fractional differential equations. Since it relies on the solution of shifted linear systems, it gives effective methods for all cases where the matrix is sparse. Our approach using matrix exponentials is instead more practical when aiming at exploiting the Kronecker structure in the operator.

4 Numerical experiments

In this last section we report a few numerical experiments that further validate our bounds, showing in which cases they are most descriptive. In addition, we show that the exponential sum expansions yield an effective solver for problems with a low-rank right-hand side.

All numerical experiments have been run on an AMD Ryzen 7 3700x CPU with 32GB of RAM, running MATLAB 2022a with the bundled Intel MKL BLAS. The code for the experiments can be found at https://github.com/numpi/fractional-expsums.

4.1 3D fractional poisson equation

as a first example, we consider the solution of the fractional Poisson equation on the 3-dimensional cube $[0, 1]^3$:

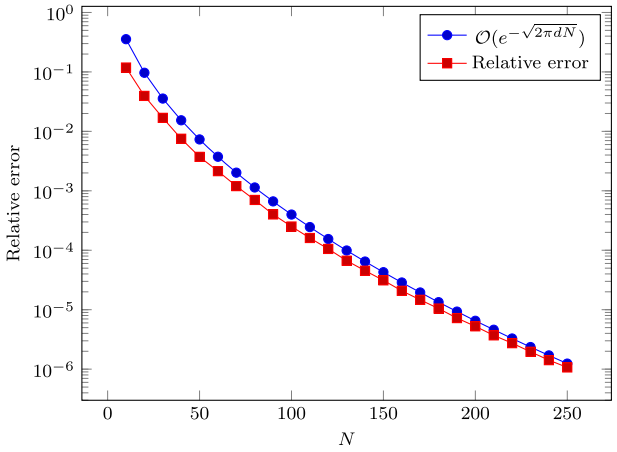

Fig. 3 Relative error on the approximation of the solution for the discretization of the problem in (15) using 128^3 points, with $\alpha = 0.4$. The error is computed using the Frobenius norm, and the approximation is computed using exponential sums with N terms, as in Theorem 2

$$\begin{cases} (-\Delta)^{\alpha} u = f & \text{in } \Omega \\ u \equiv 0 & \text{on } \partial\Omega \end{cases}, \qquad \Omega = [0,1]^3. \tag{15}$$

We discretize the domain with a uniformly spaced grid with 128 points in each direction, and the operator Δ by finite differences, which yields the linear system

$$\mathcal{A} = \bigoplus_{i=1}^{3} A_i, \qquad A_i = \frac{1}{h^2} \begin{bmatrix} 2 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix}.$$

where $h = \frac{1}{n-1}$ is the distance between the grid points. We approximate the solution of (15) by computing $\mathbf{u} = \mathcal{A}^{-\alpha}\mathbf{f}$ where $\mathbf{f}$ is the vector containing the evaluations of $f(x, y, z) = 1/(1 + x + y + z)$ at the internal points of the discretization grid. For this example, we choose $\alpha = 0.4$.

In Fig. 3 we report the quality of the approximation obtained for $\mathbf{u}$ by using the exponential sum with N terms described in Theorem 2. The exact solution is computed by diagonalizing the matrices A_i, which is feasible and accurate because they are symmetric and of moderate sizes.

We now consider the same example with right-hand side $f(x, y, z) = \sin x \cos y e^z$. Since this function is separable, the corresponding vector $\mathbf{f}$ is the vectorization of a rank 1 tensor. This allows to directly build a low-rank approximation of the solution by the expansion:

$$\mathcal{A}^{-\alpha}\mathbf{x} \approx \sum_{j=1}^{N} \alpha_j e^{-\beta_j \mathcal{A}} \mathbf{f},$$

Table 1 Time and accuracy of the low-rank approximation to $\mathcal{A}^{-\alpha}\mathbf{f}$ for $\alpha = \frac{1}{2}$ obtained by the exponential sums of length $N = 100, 200, 350$, and runtime of the dense evaluation based on diagonalization, for $d = 3$

n	t_{dense}	t_{100}	res_{100}	t_{200}	res_{200}	t_{350}	res_{350}
128	0.15	0.15	$1.26 \cdot 10^{-4}$	0.3	$1.85 \cdot 10^{-6}$	0.6	$1.62 \cdot 10^{-8}$
256	1	0.57	$1.27 \cdot 10^{-4}$	0.95	$1.86 \cdot 10^{-6}$	1.8	$1.63 \cdot 10^{-8}$
512	8.1	2.03	$1.28 \cdot 10^{-4}$	3.61	$1.87 \cdot 10^{-6}$	6.29	$1.64 \cdot 10^{-8}$
1024	-	10.1	–	20.6	–	35.15	–
2048	-	52.35	–	104.3	–	182.1	–
4096	-	290.8	–	568.6	–	926.7	–

$$\mathbf{f} = \text{vec}(F), \qquad e^{-\beta_j \mathcal{A}}\mathbf{f} = \text{vec}(F \times_1 e^{-\beta_1 A_1} \ldots \times_d e^{-\beta_d A_d}) \tag{16}$$

Under these hypotheses, the cost of evaluating the inverse fractional power is dominated by computing the matrix exponentials, and requires $\mathcal{O}(dn^3 + Ndn^2)$ flops for a d-dimensional tensor with all modes of length n. In contrast, evaluating the fractional power by diagonalization requires $\mathcal{O}(dn^4)$ flops. In Table 1 we compare the cost of these two algorithms, using a different length of the exponential sum approximation to $z^{-\alpha}$ for $\alpha = \frac{1}{2}$.

We note that in this case, it is not practical to compute the dense solution for large dimensions, since the memory required is $\mathcal{O}(n^d)$; the low-rank approximation obtained through (16) only requires $\mathcal{O}(nd)$ storage. For this reason, we only report the results for the dense case and the accuracy up to dimension $n = 512$ in Table 1.

We remark that since the convergence bound is uniform over $[1, +\infty)$ the accuracy does not degrade as $n \to \infty$, even if the largest eigenvalues of the discretized Laplacian converge to infinity; this is necessarily the case, since the underlying continuous operator is unbounded.

If n grows and the A_i are structured, it can be convenient to exploit strategies to directly compute $e^{-\beta_j A_i} v$ instead of building the entire matrix exponential $e^{-\beta_j A_i}$, such as methods based on Krylov subspaces (see [19] and the references therein) or on truncated Taylor expansions [4].

4.2 Low-rank approximability in tensor formats

To test the results concerning low-rank approximability, we solve an equation in the form $\mathcal{A}^{\alpha} x = c$, then we check the distance of the solution with the closest rank j tensor, and we compare it with the upper bound from Theorem 3. We choose as A_i the discretization of the 1D Laplacian as in Sect. 4.1, and the right-hand side c as $c = c_1 \otimes \ldots \otimes c_d$, with c_i containing entries distributed as independent Gaussian random variables with mean 0 and variance 1.

We have computed the reference solution explicitly by diagonalization of the A_i. Then, we have approximated for each $r = 1, \ldots, 40$ the best CP approximant of rank at most r using the `cp_als` algorithm in the Tensor Toolbox [21] and `cpd` from TensorLab [36], and for each N we have chosen the best approximation. The decay

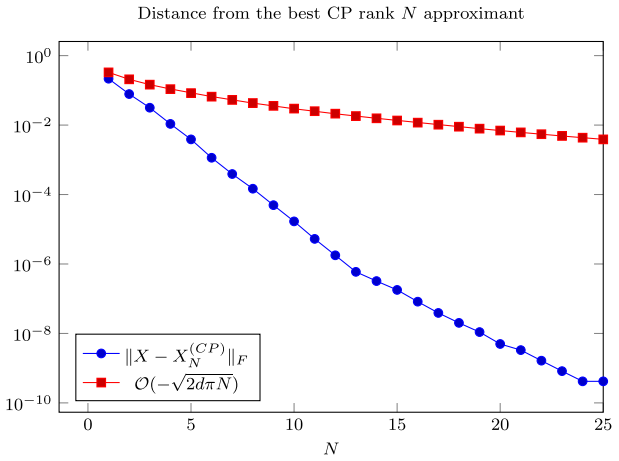

Fig. 4 Distance of the solution X from the best approximant of CP rank at most N, approximated by the best approximation obtained from `cp_als` in the Tensor Toolbox and `cpd` from TensorLab. The distance is compared with the upper bound for the asymptotic decay rate predicted by Theorem 3

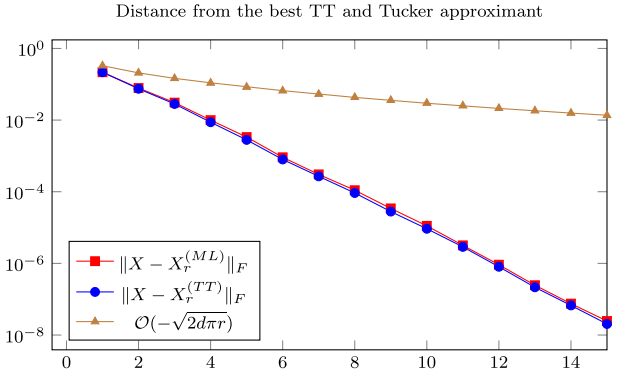

Fig. 5 Distance of the solution X from the best approximant of multilinear and TT ranks at most r, denoted by $X_r^{(ML)}$ and $X_r^{(TT)}$ and approximated by `hosvd` in the Tensor Toolbox and by the TT-Toolbox, respectively. The distance is compared with the upper bound for the asymptotic decay rate predicted by Theorem 3

rate is compared with $\mathcal{O}(e^{-\sqrt{2\pi dr}})$ predicted by Theorem 3 in Fig. 4. The problem is chosen of size $n_1 = n_2 = n_3 = 128$, the power $\alpha = \frac{1}{2}$, and the tolerance for the `cp_als` algorithm is set to 10^{-12}, and a maximum of 100 iterations. The parameters for `cpd` have not been tuned, as they were already providing good results out of the box.

The estimate turns out to be somewhat pessimistic (the convergence of low-rank approximant in CPD format is faster than what we predict), but is closer than what we will obtain in the HOSVD and TT cases.

We have then run the same tests for multilinear and Tensor-Train ranks, which are much smaller. In this context, our prediction of approximability turns out to be very pessimistic, as visible in Fig. 5.

We believe that the definition of ranks for the multilinear and TT cases, involving matricizations, may be analyzed with more powerful tools from matrix theory, and hence obtain stronger decay bounds.

The bounds are not completely descriptive of the decay rate, but can be used to justify the application of low-rank methods to the problems under consideration, since they provide easily computable a-priori bounds.

4.3 High-dimensional fractional PDEs with tensor-trains

We consider the computation of the solution for the solution of the PDE $(-\Delta)^{\alpha} = f$ over $[0, 1]^d$, with large d, and we choose the function $f(x_1, \ldots, x_d)$ as follows:

$$f(x_1, \ldots, x_d) := \frac{1}{1 + x_1 + \cdots + x_d}, \qquad x_i \in [0, 1].$$

This function has low multilinear and tensor train ranks [32], but methods based on the Tucker decomposition are not suitable, because of the exponential storage cost in d. On the other hand, the CPD of a function not directly given in a separable form is not easy to compute in general. Hence, we focus on solving the equation in a Tensor-Train format.

As we did in Sect. 4.1, we discretize the domain with a uniformly spaced grid with 128 points in each direction and we compute $\mathbf{u} = \mathcal{A}^{-\alpha}\mathbf{f}$, where $\mathcal{A}$ is the discretization of $-\Delta$ and $\mathbf{f}$ is the vector containing all the evaluations of f at the internal points of the discretization grid.

To obtain a Tensor-Train representation of $\mathbf{f}$, the tensor with the evaluations of $f(x_1, \ldots, x_d)$ at the grid points, we relied on an AMEn-based version of the TT-cross approximation, as described in [13], and implemented in the TT-Toolbox. This only requires to evaluate $f(x_1, \ldots, x_d)$ at a few specific points in the grid, making the method very effective in practice.

We then use our exponential sum approximation with N terms, which requires to compute the Nd matrix exponentials $e^{-\beta_j A_i}$ for $j = 1, \ldots, N$ and $i = 1, \ldots, d$, and then to multiply them by a low TT rank matrix with a mode-j product. The latter can be evaluated efficiently in the Tensor-Train arithmetic, and the storage for the result of the partial sum is kept under control by recompressing the tensors with economy TT-SVDs, as implemented in the `round` command of the TT-Toolbox [28].

In Table 2 we report time and accuracy for the approximation of $\mathbf{u}$ by the exponential sum with $N = 200$ terms for $\alpha = \frac{1}{2}$. Moreover, we report the TT-rank of the approximated solution.

5 Conclusions and outlook

We have developed an exponential sum approximation for $z^{-\alpha}$, with $0 < \alpha < 1$; this finds application in solving linear systems involving fractional powers of Kronecker sums, that naturally appear when treating high-dimensional fractional diffusion problems on tensorized domains using the spectral fractional Laplacian.

Table 2 Time, accuracy and rank of the final solution, of the low-rank approximation to $\mathcal{A}^{-\alpha}\mathbf{f}$ for $\alpha = \frac{1}{2}$ on $[0, 1]^d$ obtained by the exponential sums of length $N = 200$, and different choices of d

d	Time(s)	Error	Rank
2	0.42	$1.65 \cdot 10^{-6}$	15
3	0.88	$1.76 \cdot 10^{-6}$	16
4	1.87	$1.87 \cdot 10^{-6}$	24
6	4.22	–	26
10	9.82	–	28
15	16.2	–	27
20	24.5	–	27

The accuracy is computed by comparing the approximated solution with the one obtained by solving the Sylvester equation by diagonalization, which is only feasible for small $d \leq 4$

The design of this explicit approximation (along with guaranteed error bounds) allows to effectively solve such linear systems for generic right-hand sides, but is particularly interesting when the right-hand side is stored in a low-rank tensor format. For these cases (examples have been reported for CP, Tensor-Trains, and Tucker tensors) the exponential sum yields an explicit approximate solution in the same low-rank tensor format, exploiting the preservation of the rank under mode-j products.

An important consequence of the construction is to predict the approximability of the solution in the same format of the right-hand sides. We have tested the quality of such predictions, and we have verified that it is not completely descriptive of the approximation speed in the CP format and for TT and multilinear ranks. The relatively loose upper bound derived may depend on the fact that we are developing an approximant that is accurate over an unbounded interval $[1, \infty)$ whereas the spectrum of discretized differential operators is limited. On the other hand, our theory gives dimension-independent approximability bounds — which is clearly visible in our experiments, where the approximation error arising from the low-rank truncation is always controlled a-priori when the number of terms in the exponential sum is fixed, and is not influenced by the number of discretization points. Our results can be extended to provide a priori justification for the approximability for infinite dimensional operators with unbounded spectra, and in general motivate the use of adaptive rank truncations.

We believe that other tools may be used to extend our bounds to the TT and the Tucker case by restricting the focus to bounded spectra; this will be investigated in future work. Giving up the generality of unbounded operators is likely to allow for a more realistic description of the problem.

Since the latter formats (TT and Tucker) allow for easy recompressions, the proposed exponential sum approach can be a competitive solver even if the ranks in the solution are slightly overestimated. This has been demonstrated in a few practical cases. In particular, in the TT case this framework allows to treat very high-dimensional problems, beating the curse-of-dimensionality.

Acknowledgements Partial financial support was received through an INdAM/GNCS Project "Metodi low-rank per problemi di algebra lineare con struttura data-sparse".

Funding Open access funding provided by Università di Pisa within the CRUI-CARE Agreement.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Aceto, L., Novati, P.: Rational approximation to the fractional Laplacian operator in reaction-diffusion problems. SIAM J. Sci. Comput. **39**(1), A214–A228 (2017)
2. Aceto, L., Novati, P.: Rational approximations to fractional powers of self-adjoint positive operators. Numer. Math. **143**(1), 1–16 (2019)
3. Aceto, L., Novati, P.: Exponentially convergent trapezoidal rules to approximate fractional powers of operators. J. Sci. Comput. **91**(2), 1–18 (2022)
4. Al-Mohy, A.H., Higham, N.J.: Computing the action of the matrix exponential, with an application to exponential integrators. SIAM J. Sci. Comput. **33**(2), 488–511 (2011)
5. Beckermann, B., Townsend, A.: On the singular values of matrices with displacement structure. SIAM J. Mat. Anal. Appl. **38**(4), 1227–1248 (2017)
6. Benner, P., Li, R.-C., Truhar, N.: On the ADI method for Sylvester equations. J. Comput. Appl. Math. **233**(4), 1035–1045 (2009)
7. Benzi, M., Simoncini, V.: Approximation of functions of large matrices with Kronecker structure. Numer. Math. **135**(1), 1–26 (2017)
8. Bini, D.A., Massei, S., Robol, L.: On the decay of the off-diagonal singular values in cyclic reduction. Linear Algebra Appl. **519**, 27–53 (2017)
9. Bonito, A., Pasciak, J.: Numerical approximation of fractional powers of elliptic operators. Math. Comp. **84**(295), 2083–2110 (2015)
10. Börm, S., Grasedyck, L., Hackbusch, W.: Introduction to hierarchical matrices with applications. Eng. Anal. Bound. Elem. **27**(5), 405–422 (2003)
11. Cirac, J.I., Verstraete, F.: Renormalization and tensor product states in spin chains and lattices. J. Phys. A **42**(50), 504004 (2009)
12. De Lathauwer, L., De Moor, B., Vandewalle, J.: A multilinear singular value decomposition. SIAM J. Matrix Anal. Appl. **21**(4), 1253–1278 (2000)
13. Dolgov, S.V., Savostyanov, D.V.: Alternating minimal energy methods for linear systems in higher dimensions. SIAM J. Sci. Comput. **36**(5), A2248–A2271 (2014)
14. Grasedyck, L.: Hierarchical singular value decomposition of tensors. SIAM J. Matrix Anal. Appl. **31**(4), 2029–2054 (2010)
15. Grasedyck, L., Kressner, D., Tobler, C.: A literature survey of low-rank tensor approximation techniques. GAMM-Mitt. **36**(1), 53–78 (2013)
16. Hackbusch, W.: Hierarchical matrices: algorithms and analysis, vol. 49, Springer (2015)
17. Harizanov, S., Lazarov, R., Margenov, S.: A survey on numerical methods for spectral space-fractional diffusion problems. Fract. Calc. Appl. Anal. **23**(6), 1605–1646 (2020)
18. Harizanov, S., Lazarov, R., Margenov, S., Marinov, P., Pasciak, J.: Analysis of numerical methods for spectral fractional elliptic equations based on the best uniform rational approximation. J. Comput. Phys. **408**, 109285 (2020)
19. Higham, N.J.: Functions of matrices: theory and computation. SIAM (2008)

20. Ilic, M., Liu, F., Turner, I., Anh, V.: Numerical approximation of a fractional-in-space diffusion equation. I. Fract. Calc. Appl. Anal. **8**(3), 323p–341p (2005)
21. Kolda, T.G., Bader, B.W.: Tensor decompositions and applications. SIAM Rev. **51**(3), 455–500 (2009)
22. Kressner, D.: Bivariate matrix functions. Tech. Rep. (2011)
23. Kressner, D.: A Krylov subspace method for the approximation of bivariate matrix functions. In: Structured matrices in numerical linear algebra, pp. 197–214. Springer (2019)
24. Kressner, D., Tobler, C.: Krylov subspace methods for linear systems with tensor product structure. SIAM J. Matrix Anal. Appl. **31**(4), 1688–1714 (2010)
25. Massei, S., Mazza, M., Robol, L.: Fast solvers for two-dimensional fractional diffusion equations using rank structured matrices. SIAM J. Sci. Comput. **41**(4), A2627–A2656 (2019)
26. Massei, S., Robol, L.: Rational Krylov for Stieltjes matrix functions: convergence and pole selection. BIT **61**(1), 237–273 (2021)
27. Massei, S., Robol, L.: Mixed precision recursive block diagonalization for bivariate functions of matrices. SIAM J. Matrix Anal. Appl. **43**(2), 638–660 (2022)
28. Oseledets, I.V.: Tensor-train decomposition. SIAM J. Sci. Comput. **33**(5), 2295–2317 (2011)
29. Oseledets, I.V., Tyrtyshnikov, E.E.: Breaking the curse of dimensionality, or how to use SVD in many dimensions. SIAM J. Sci. Comput. **31**(5), 3744–3759 (2009)
30. Palitta, D., Simoncini, V.: Matrix-equation-based strategies for convection-diffusion equations. BIT **56**(2), 751–776 (2016)
31. Schollwöck, U.: The density-matrix renormalization group in the age of matrix product states. Ann. Phys. **326**(1), 96–192 (2011)
32. Shi, T., Townsend, A.: On the compressibility of tensors. SIAM J. Matrix Anal. Appl. **42**(1), 275–298 (2021)
33. Simoncini, V.: Computational methods for linear matrix equations. SIAM Rev. **58**(3), 377–441 (2016)
34. Stenger, F.: Numerical methods based on Sinc and analytic functions, vol. 20, Springer Science & Business Media (2012)
35. Townsend, A.: Computing with functions in two dimensions. In: PhD Thesis, Oxford (2014)
36. Vervliet, N., Debals, O., Sorber, L., Van Barel, M., De Lathauwer, L.: Tensorlab 3.0, Mar. Available online (2016)
37. Yang, Q., Turner, I., Liu, F., Ilić, M.: Novel numerical methods for solving the time-space fractional diffusion equation in two dimensions. SIAM J. Sci. Comput. **33**(3), 1159–1180 (2011)
38. Zolotarev, E.I.: Application of elliptic functions to questions of functions deviating least and most from zero. Zap. Imp. Akad. Nauk. St. Petersb. **30**(5), 1–59 (1877)