Numerische Mathematik (2019) 143:1–16
https://doi.org/10.1007/s00211-019-01048-4

Numerische Mathematik

Rational approximations to fractional powers of self-adjoint positive operators

Lidia Aceto[1] · Paolo Novati[2]

Received: 18 June 2018 / Revised: 18 December 2018 / Published online: 11 May 2019

Abstract
We investigate the rational approximation of fractional powers of unbounded positive operators attainable with a specific integral representation of the operator function. We provide accurate error bounds by exploiting classical results in approximation theory involving Padé approximants. The analysis improves some existing results and the numerical experiments proves its accuracy.

Mathematics Subject Classification 47A58 · 65F60 · 65D32

1 Introduction

Let $\mathcal{H}$ be a separable Hilbert space endowed with an inner product $\langle \cdot, \cdot \rangle$ and corresponding norm $\|x\|_{\mathcal{H}} = \langle x, x \rangle^{1/2}$. Let $\mathcal{L}$ be a self-adjoint positive operator with spectrum $\sigma(\mathcal{L}) \subseteq [c, +\infty)$, $c > 0$. Moreover, assume that $\mathcal{L}$ has compact inverse. This paper deals with the numerical approximation of $\mathcal{L}^{-\alpha}$, $0 < \alpha < 1$, that, in this setting, can be defined through the spectral decomposition, i.e.,

This work was partially supported by GNCS-INdAM, University of Pisa (Grant PRA_2017_05) and FRA-University of Trieste.
The authors are members of the INdAM research group GNCS.

✉ Lidia Aceto
lidia.aceto@unipi.it

Paolo Novati
novati@units.it

[1] Dipartimento di Matematica, Università di Pisa, Via F. Buonarroti, 1/C, I-56127 Pisa, Italy

[2] Dipartimento di Matematica e Geoscienze, Università di Trieste, via Valerio 12/1, I-34127 Trieste, Italy

$$\mathcal{L}^{-\alpha}u = \sum_{s=1}^{\infty} \mu_s^{-\alpha} \langle u, \varphi_s \rangle \varphi_s,$$

where $\{\varphi_s\}_{s=1}^{\infty}$ is the orthonormal system of eigenfunctions of $\mathcal{L}$ and $\{\mu_s\}_{s=1}^{\infty}$ is the corresponding sequence of positive real eigenvalues (arranged in order of increasing magnitude and counted according to their multiplicities). Clearly $\mathcal{L}^{-\alpha}$ is a self-adjoint compact operator on $\mathcal{H}$. Since the function $\lambda^{-\alpha}$ is continuous in $[\mu_1, +\infty)$, we have that (see e.g. [22, Theorem 1.7.7])

$$\left\| \mathcal{L}^{-\alpha} \right\|_{\mathcal{H}\to\mathcal{H}} = \sup_{\lambda\in\sigma(\mathcal{L})} \left|\lambda^{-\alpha}\right| = \mu_1^{-\alpha},$$

where $\|\cdot\|_{\mathcal{H}\to\mathcal{H}}$ denotes the operator norm induced by $\|\cdot\|_{\mathcal{H}}$.

An important and widely studied example comes from certain fractional models involving the symmetric space fractional derivative $(-\Delta)^{\beta/2}$ of order β $(1 < \beta \leq 2)$ [17]; in this situation the fractional power is generally approximated through the approximation of $(-\Delta)^{\beta/2-1}$ [16].

A standard approach to approximate $\mathcal{L}^{-\alpha}$ is by means of $\mathcal{L}_N^{-\alpha}$ where $\mathcal{L}_N$ is a finite dimensional self-adjoint positive operator representing a discretization of $\mathcal{L}$. Clearly, improving the sharpness of the discretization the typical situation is that $\lambda_{\min}(\mathcal{L}_N) \to \mu_1$ and $\lambda_{\max}(\mathcal{L}_N) \to +\infty$ ($\lambda_{\min}(\mathcal{L}_N)$ and $\lambda_{\max}(\mathcal{L}_N)$ denoting the smallest and the largest eigenvalues of $\mathcal{L}_N$).

In this framework, in order to compute $\mathcal{L}_N^{-\alpha}$ it is quite natural to employ rational forms. For instance, in [14,15] some rational approximations are obtained by considering the best uniform rational approximation of $\lambda^{1-\alpha}$ and λ^{α} on the interval [0, 1]. Beside, other well established techniques are the ones based on existing integral representations of the Markov function $\lambda^{-\alpha}$ and then on the use of suitable quadrature rules that finally lead to rational approximations of the type

$$\mathcal{L}_N^{-\alpha} \approx \mathcal{R}_{k-1,k}(\mathcal{L}_N), \quad \mathcal{R}_{k-1,k}(\lambda) = \frac{p_{k-1}(\lambda)}{q_k(\lambda)}, \quad p_{k-1} \in \Pi_{k-1},\ q_k \in \Pi_k,$$

where Π_j denotes the set of polynomials of degree j (see e.g. [5,11,20]).

In this setting, in [1–3] the rational forms arise from the use of the Gauss–Jacobi rule for computing the integral representation (see [4, Eq. (V.4) p. 116])

$$\mathcal{L}^{-\alpha} = \frac{\sin(\alpha\pi)}{(1-\alpha)\pi} \int_0^{\infty} (\rho^{1/(1-\alpha)} I + \mathcal{L})^{-1} d\rho, \tag{1}$$

after the change of variable

$$\rho^{1/(1-\alpha)} = \tau \frac{1-t}{1+t}, \qquad \tau > 0. \tag{2}$$

Working in finite dimension, the asymptotically optimal choice of the parameter τ, yields an error of type

$$\mathcal{O}\left(\exp\left(-4k\sqrt[4]{\lambda_{\min}(\mathcal{L}_N)/\lambda_{\max}(\mathcal{L}_N)}\right)\right), \tag{3}$$

where k is the number of points of the quadrature rule, corresponding to a $\mathcal{R}_{k-1,k}(\lambda)$ rational form. Of course $\sqrt[4]{\lambda_{\min}(\mathcal{L}_N)/\lambda_{\max}(\mathcal{L}_N)} \to 0$ improving the quality of the discretization so that (3) becomes meaningless whenever $\mathcal{L}_N$ represents an arbitrarily sharp discretization of $\mathcal{L}$.

The basic aim of the present work is to overcome this problem by working in the infinite dimensional setting. Using the fact that the Gauss–Jacobi quadrature on Markov functions is related to the Padé approximation, we derive an expression for the truncation error $\lambda^{-\alpha} - \mathcal{R}_{k-1,k}(\lambda) := \lambda^{-\alpha} - \tau^{-\alpha} R_{k-1,k}(\lambda/\tau)$ (here $R_{k-1,k}(\lambda/\tau)$ denotes the $(k-1,k)$-Padé approximant of $(\lambda/\tau)^{-\alpha}$), that leads to an alternative definition of the parameter τ independent of the discretization and, at the same time, ensuring an asymptotically optimal rate of convergence. In particular, we are able to show that the quadrature nodes for (1) can be defined so that the error for the computation of $\mathcal{L}^{-\alpha}$ decays approximatively like

$$\left\| \mathcal{L}^{-\alpha} - \tau^{-\alpha} R_{k-1,k}\left(\frac{\mathcal{L}}{\tau}\right) \right\|_{\mathcal{H}\to\mathcal{H}} \approx \sin(\alpha\pi) c^{-\alpha} \left(\frac{2ke^{1/2}}{\alpha}\right)^{-4\alpha},$$

and therefore sublinearly. Qualitatively, a similar behavior can also been observed by working with rational Krylov methods to approximate the action of functions involving $\mathcal{L}^{-\alpha}$ (see, e.g., [19]), in which the error decays like k^{-p}, where $p > 0$ depends on the function. The sublinearity appears when considering unbounded spectra. Using the analysis for unbounded operators, we also show how to improve quantitatively (3) whenever we assume to work with $\mathcal{L}_N$. The key point consists in taking τ in (2) dependent on k.

We remark that all the theory here developed can be easily employed to compute the action of the unbounded operator $\mathcal{L}^{1-\alpha}$ on a vector $f \in D(\mathcal{L})$ ($D(\mathcal{L})$ is the domain of $\mathcal{L}$), that is $\mathcal{L}^{1-\alpha} f$. This may occur for instance when solving equations involving the above mentioned fractional Laplacian. In this situation, after evaluating $g = \mathcal{L} f$, $\mathcal{L}^{1-\alpha} f$ can be computed using our analysis on $\mathcal{L}^{-\alpha} g$. Nevertheless, the poles of the rational forms here derived can also be used to compute $\mathcal{L}^{-\alpha} g$ by means of a rational Krylov method.

Finally, since the subject of this paper is closely related to matrix pth roots, we mention here [6,8,9,21,23] in which other approaches such as the Newton method were developed.

The paper is organized as follows. In Sect. 2 we recall the basic features of the Gauss–Jacobi based rational forms for computing (1). Section 3 contains the error analysis and represents the main contribution of this paper. In Sect. 4 we revisit the error analysis for the case of bounded spectra. Finally, in Sect. 5 we present some numerical experiments that validate the theoretical results.

2 Background on the Gauss–Jacobi approach

Starting from the representation (1), in order to approximate the fractional Laplacian in [2] the authors consider the change of variable (2), that leads to

$$\mathcal{L}^{-\alpha} = \frac{2\sin(\alpha\pi)\tau^{1-\alpha}}{\pi} \int_{-1}^{1} (1-t)^{-\alpha} (1+t)^{\alpha-2} \left(\tau \frac{1-t}{1+t} I + \mathcal{L} \right)^{-1} dt. \quad (4)$$

Using the k-point Gauss–Jacobi rule with respect to the weight function $\omega(t) = (1-t)^{-\alpha} (1+t)^{\alpha-1}$ the above integral is approximated by the rational form

$$\mathcal{L}^{-\alpha} \approx \sum_{j=1}^{k} \gamma_j (\eta_j I + \mathcal{L})^{-1} := \tau^{-\alpha} R_{k-1,k} \left(\frac{\mathcal{L}}{\tau} \right), \quad (5)$$

where the coefficients γ_j and η_j are given by

$$\gamma_j = \frac{2\sin(\alpha\pi)\tau^{1-\alpha}}{\pi} \frac{w_j}{1+\vartheta_j}, \qquad \eta_j = \frac{\tau(1-\vartheta_j)}{1+\vartheta_j}; \quad (6)$$

here w_j and ϑ_j are, respectively, the weights and nodes of the Gauss–Jacobi quadrature rule.

The choice of τ in (2) is crucial for the quality of the approximation attainable by (5). As already mentioned in the Introduction, working with bounded operators, it has been shown in [2] that asymptotically, that is for $k \to +\infty$, the optimal choice is given by

$$\widetilde{\tau} = \sqrt{\lambda_{\min}(\mathcal{L}_N)\lambda_{\max}(\mathcal{L}_N)}. \quad (7)$$

With this choice and denoting by $\kappa(\mathcal{L}_N)$ the spectral condition number of $\mathcal{L}_N$ and by $\|\cdot\|_2$ the induced Euclidean norm, we obtain

$$\left\| \mathcal{L}_N{}^{-\alpha} - \widetilde{\tau}^{-\alpha} R_{k-1,k} \left(\frac{\mathcal{L}_N}{\widetilde{\tau}} \right) \right\|_2 \leq C \left(\frac{\sqrt[4]{\kappa(\mathcal{L}_N)} - 1}{\sqrt[4]{\kappa(\mathcal{L}_N)} + 1} \right)^{2k}, \quad (8)$$

with C independent of k, which is a sharper version of (3). We remark that $\widetilde{\tau}$ is independent of k. In what follows we shall follow a different strategy allowing a dependence on k (in any case the coefficients γ_j and η_j completely change with k) but at the same time a 'mesh-independence', since we work with the unbounded operator $\mathcal{L}$.

3 Error analysis

Working with the ratio λ/τ, where $\lambda \in [c, +\infty)$ and $\tau > 0$, as shown in [12, Lemma 4.4] the k-point Gauss–Jacobi quadrature given by (5)–(6) is such that $R_{k-1,k}(\lambda/\tau)$

corresponds to the $(k-1,k)$-Padé approximant of $(\lambda/\tau)^{-\alpha}$ centered at 1. In this sense, defining

$$z = 1 - \frac{\lambda}{\tau}, \tag{9}$$

in what follows we focus the attention on the $(k-1,k)$-Padé approximation

$$(1-z)^{-\alpha} \approx R_{k-1,k}(1-z).$$

Indicating the truncation error by

$$E_{k-1,k}(1-z) := (1-z)^{-\alpha} - R_{k-1,k}(1-z), \tag{10}$$

we have the following result.

Theorem 1 *For each integer $k \geq 1$ and $|arg\ (1-z)| < \pi$ the exact representation of the truncation error defined by (10) is given by*

$$E_{k-1,k}(1-z) = \frac{\Gamma(k+1-\alpha)\Gamma(k+1)}{\Gamma(1-\alpha)\Gamma(2k+1)} \frac{{}_2F_1(k+1,k+\alpha;2k+1;z)}{{}_2F_1(-k,k;\alpha;z^{-1})}(-z)^k, \tag{11}$$

in which Γ denotes the gamma function and ${}_2F_1$ the hypergeometric function.

Proof Since [18, Eq. (9.8.1)]

$$(1-z)^{-\alpha} = {}_2F_1(1,\alpha;1;z), \qquad |\arg\ (1-z)| < \pi,$$

the expression for the truncation error is obtained following the analysis given in [7, Sect. 3]. □

Proposition 1 *For $z < 1$, let $v = 1 - 2z^{-1}$ and ξ be defined by*

$$v \pm \left(v^2-1\right)^{1/2} = e^{\pm\xi}. \tag{12}$$

Then, for large values of k we have

$$E_{k-1,k}(1-z) = 4\sin(\alpha\pi)\frac{v-1}{e^{(2k+1)\xi}} \frac{\left(1+e^{-\xi}\right)^{-2\alpha}}{\left(1-e^{-\xi}\right)^{2(1-\alpha)}}\left(1+\mathcal{O}\left(\frac{1}{k}\right)\right). \tag{13}$$

Proof Since $z = 2/(1-v)$, using [10, Eqs. (16) and (17) p. 77] we have that

$$\begin{aligned}{}_2F_1(k+1,k+\alpha;2k+1;z) &= 4\frac{\Gamma(2k+1)\Gamma(1/2)}{\Gamma(k+\alpha)\Gamma(k+1-\alpha)}\frac{k^{-1/2}}{(-z)^{k+1}}\\ &\times e^{-(k+1)\xi}(1-e^{-\xi})^{-3/2+\alpha}(1+e^{-\xi})^{-1/2-\alpha}\left(1+\mathcal{O}\left(1/k\right)\right),\end{aligned}$$

$$
\begin{aligned}
{}_2F_1(-k,k;\alpha;z^{-1}) &= \frac{\Gamma(k+1)\Gamma(\alpha)}{2\Gamma(1/2)\Gamma(k+\alpha)}k^{-1/2} \\
&\quad\times (1-e^{-\xi})^{1/2-\alpha}(1+e^{-\xi})^{\alpha-1/2}\left(e^{k\xi}+e^{\pm i\pi(\alpha-1/2)}e^{-k\xi}\right)(1+\mathcal{O}(1/k)).
\end{aligned}
$$

Plugging these relations in (11) and using the identities $\Gamma(1/2) = \sqrt{\pi}$ and $\Gamma(\alpha)\Gamma(1-\alpha) = \pi/\sin(\pi\alpha)$, we find the result. □

Remark 1 As pointed out in [7, p. 402], it can be observed that eq. (13) provides a very good estimate of $E_{k-1,k}(1-z)$ even for small values of k.

Proposition 2 *For large values of k, the following representation for the truncation error holds*

$$
E_{k-1,k}\left(\frac{\lambda}{\tau}\right) = 2\sin(\alpha\pi)\left(\frac{\lambda}{\tau}\right)^{-\alpha}\left[\frac{\lambda^{1/2}-\tau^{1/2}}{\lambda^{1/2}+\tau^{1/2}}\right]^{2k}\left(1+\mathcal{O}\left(\frac{1}{k}\right)\right). \tag{14}
$$

Proof Using (12), after some algebra we obtain

$$
2\frac{v-1}{e^{(2k+1)\xi}}\frac{\left(1+e^{-\xi}\right)^{-2\alpha}}{\left(1-e^{-\xi}\right)^{2(1-\alpha)}} = \left(\frac{v+1}{v-1}\right)^{-\alpha}\frac{1}{\left[v+\left(v^2-1\right)^{1/2}\right]^{2k}}. \tag{15}
$$

Since $z = 2/(1-v)$ and $z = 1-\lambda/\tau$ we find

$$
v = \frac{\lambda+\tau}{\lambda-\tau}.
$$

Substituting this expression in (15) and then the result in (13), we easily obtain the statement. □

By (5), (9) and (10) we have

$$
\left\|\mathcal{L}^{-\alpha} - \tau^{-\alpha}R_{k-1,k}\left(\frac{\mathcal{L}}{\tau}\right)\right\|_{\mathcal{H}\to\mathcal{H}} \le \max_{\lambda\ge c}\tau^{-\alpha}\left|E_{k-1,k}\left(\frac{\lambda}{\tau}\right)\right|. \tag{16}
$$

As consequence, a suitable value for τ can be found by working with (14). To this purpose, let us consider the function

$$
f(\lambda,\tau) := \left(\frac{\lambda}{\tau}\right)^{-\alpha}\left[\frac{\lambda^{1/2}-\tau^{1/2}}{\lambda^{1/2}+\tau^{1/2}}\right]^{2k}, \tag{17}
$$

which is the τ-dependent factor of (14). We want to solve

$$
\min_{\tau>0}\max_{\lambda\ge c}\tau^{-\alpha}f(\lambda,\tau). \tag{18}
$$

For any fixed $\tau > 0$, $f(\lambda, \tau) \to +\infty$ for $\lambda \to 0^+$, $f(\lambda, \tau) \to 0$ for $\lambda \to +\infty$, $f(\lambda, \tau) = 0$ for $\lambda = \tau$ (the minimum) and, by solving $\frac{\partial f(\lambda,\tau)}{\partial \lambda} = 0$, we find a maximum at

$$\overline{\lambda} = \frac{\left(k + \sqrt{k^2+1}\right)^2}{\alpha^2}\tau = s_k^2 \frac{4k^2}{\alpha^2}\tau, \tag{19}$$

where

$$1 < s_k^2 = 1 + \frac{1}{2k^2} + \mathcal{O}(1/k^4). \tag{20}$$

Clearly $\overline{\lambda} > \tau$ and hence

$$\max_{\lambda \geq c} \tau^{-\alpha} f(\lambda, \tau) = \max\left\{\tau^{-\alpha} f(c, \tau), \tau^{-\alpha} f(\overline{\lambda}, \tau)\right\}.$$

Setting

$$\varphi_1(\tau) := \tau^{-\alpha} f(c, \tau), \quad \varphi_2(\tau) := \tau^{-\alpha} f(\overline{\lambda}, \tau), \tag{21}$$

by (17) we find

$$\begin{aligned}\varphi_1(\tau) &= c^{-\alpha}\left[\frac{c^{1/2} - \tau^{1/2}}{c^{1/2} + \tau^{1/2}}\right]^{2k},\\ \varphi_2(\tau) &= \tau^{-\alpha} f\left(s_k^2 \frac{4k^2}{\alpha^2}\tau, \tau\right)\\ &= \tau^{-\alpha}\left(s_k^2 \frac{4k^2}{\alpha^2}\right)^{-\alpha}\left(\frac{2ks_k - \alpha}{2ks_k + \alpha}\right)^{2k}\\ &= \tau^{-\alpha}\left(\frac{4k^2e^2}{\alpha^2}\right)^{-\alpha}(1 + \mathcal{O}(1/k^2)),\end{aligned} \tag{22}$$

where the last equality follows from (20) and by considering the Taylor expansion around $y = 0$ after setting $s_k^2 = 1 + y$. Since $\varphi_2(\tau)$ is monotone decreasing, whereas $\varphi_1(\tau)$ is monotone increasing for $\tau > c$, the solution of (18) is obtained by solving

$$\varphi_1(\tau) = \varphi_2(\tau) \text{ for } \tau > c. \tag{23}$$

Proposition 3 *Let τ^* be the solution of (23). Then, for k large enough,*

$$\tau^* \approx \tau_k := c\left(\frac{\alpha}{2ke}\right)^2 \exp\left(2W\left(\frac{4k^2e}{\alpha^2}\right)\right), \tag{24}$$

where W denotes the Lambert-W function.

Proof Neglecting the factor $(1 + \mathcal{O}(1/k^2))$ in (22), Eq. (23) implies

$$\left(\frac{c}{\tau}\right)^{-\alpha}\left[\frac{\tau^{1/2} - c^{1/2}}{\tau^{1/2} + c^{1/2}}\right]^{2k} = \left(\frac{4k^2e^2}{\alpha^2}\right)^{-\alpha}. \tag{25}$$

Setting $x = (c/\tau)^{1/2} < 1$ and

$$a_k = \frac{\alpha}{2ke}, \tag{26}$$

by (25) we obtain

$$x^{-\frac{\alpha}{k}} \left(\frac{1-x}{1+x} \right) = a_k^{\frac{\alpha}{k}}.$$

Since $(1+x)^{-1} = 1 - x + \mathcal{O}(x^2)$, using the approximation

$$\frac{1-x}{1+x} \approx e^{-2x} \tag{27}$$

we solve

$$e^{-2x} = (a_k x)^{\frac{\alpha}{k}}.$$

Therefore

$$-2x = \frac{\alpha}{k} \ln (a_k x)$$

which implies

$$\frac{2k}{a_k \alpha} = \frac{1}{a_k x} \ln \left(\frac{1}{a_k x} \right).$$

Using the Lambert-W function, the solution for such equation is given by

$$\frac{1}{a_k x} = \exp \left(W \left(\frac{2k}{a_k \alpha} \right) \right).$$

Substituting x by $(c/\tau)^{1/2}$ and using (26) we obtain the expression of τ_k. □

In order to appreciate the approximation given by (24), working with $\alpha = 0.6$ and $c = 1$, in Fig. 1 we plot τ^* and τ_k for small values of k, on the left, and their relative distance for $k = 1, 2, \ldots, 200$ on the right. Moving α or c we obtain similar pictures.

We remark that since for large z

$$W(z) = \ln z - \ln (\ln z) + \mathcal{O}(1),$$

we have (see (24))

$$\tau_k = c \frac{4k^2}{\alpha^2} \left[\ln \left(\frac{4k^2}{\alpha^2} e \right) \right]^{-2} (1 + \mathcal{O}(1/k^2)). \tag{28}$$

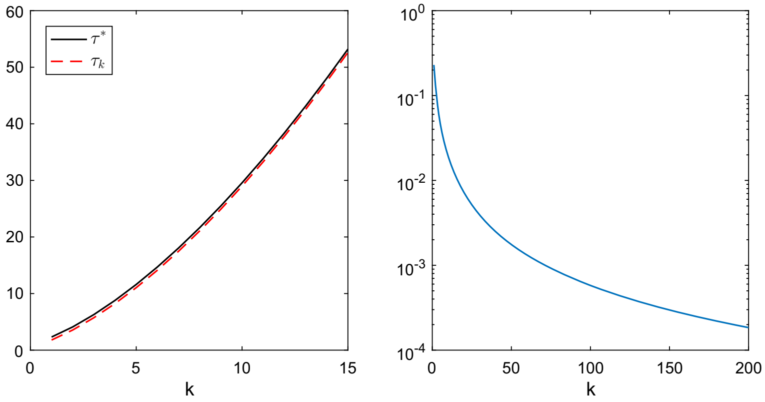

Fig. 1 On the left: for small values of k, comparison between τ^*, the exact solution of (23) (numerically evaluated), and τ_k as defined by (24). On the right, the relative distance in logarithmic scale, that is, $\log_{10}\left(\left|\frac{\tau^*-\tau_k}{\tau^*}\right|\right)$, for $k = 1, 2, \ldots, 200$. In both pictures, $\alpha = 0.6$ and $c = 1$

By (22) we thus obtain

$$\begin{aligned}\varphi_2\left(\tau_k\right) &= \tau_k^{-\alpha} f(\overline{\lambda}, \tau_k) \\ &= c^{-\alpha}\left(\frac{2ke^{1/2}}{\alpha}\right)^{-4\alpha}\left[2\ln\left(\frac{2k}{\alpha}\right)+1\right]^{2\alpha}(1+\mathcal{O}(1/k^2)).\end{aligned}$$

The above analysis yields the following result.

Theorem 2 *Let τ_k be defined according to (24). Taking $\tau = \tau_k$ in (2), for k large enough we have*

$$\begin{aligned}\left\|\mathcal{L}^{-\alpha} - \tau_k^{-\alpha} R_{k-1,k}\left(\frac{\mathcal{L}}{\tau_k}\right)\right\|_{\mathcal{H}\to\mathcal{H}} &\le 2\sin(\alpha\pi)\, c^{-\alpha}\left(\frac{2ke^{1/2}}{\alpha}\right)^{-4\alpha} \\ &\times\left[2\ln\left(\frac{2k}{\alpha}\right)+1\right]^{2\alpha}\left(1+\mathcal{O}\left(\frac{1}{k^2}\right)\right). \quad (29)\end{aligned}$$

Proof The statement immediately follows from (14), (16), and the analysis just made. □

Remark 2 The factor $c^{-\alpha}$ in the bound (29) reveals how the problem becomes increasingly difficult if the spectrum is close to the branch point of $\lambda^{-\alpha}$.

4 The case of bounded operators

The theory just developed can be easily adapted to the case of bounded operators $\mathcal{L}_N$ with spectrum contained in $[c, \lambda_N]$, where $\lambda_N = \lambda_{\max}(\mathcal{L}_N)$. In this situation we want to solve

$$\min_{\tau>0} \max_{c\leq\lambda\leq\lambda_N} \tau^{-\alpha} f(\lambda,\tau). \tag{30}$$

Looking at (19) we have $\overline{\lambda} = \overline{\lambda}(k) \to +\infty$ as $k \to +\infty$. As a consequence, for $\overline{\lambda} \leq \lambda_N$ (k small), the solution of (30) remains the one approximated by (24) and the bound (29) is still valid. On the contrary, for $\overline{\lambda} > \lambda_N$ (k large), the bound can be improved as follows.

Remembering the features of the function $f(\lambda,\tau)$ introduced in (17), we have that for $\overline{\lambda} > \lambda_N$ the solution of (30) is obtained by solving

$$\varphi_1(\tau) = \varphi_3(\tau) \text{ for } \tau > c, \tag{31}$$

where $\varphi_1(\tau)$ is defined in (21) and

$$\varphi_3(\tau) := \tau^{-\alpha} f(\lambda_N,\tau) = \lambda_N^{-\alpha}\left[\frac{\lambda_N^{1/2} - \tau^{1/2}}{\lambda_N^{1/2} + \tau^{1/2}}\right]^{2k}.$$

It can be easily verified that the equation $\varphi_1(\tau) = \varphi_3(\tau)$ has in fact two solutions, one in the interval $(0, c)$ and the other in (c, λ_N). Anyway since $\varphi_3(\tau)$ is monotone decreasing in $[0, \lambda_N)$ we have to look for the one in (c, λ_N) as stated in (31).

Proposition 4 *Let $\hat{\tau}^*$ be the solution of* (31). *Then, for k large enough,*

$$\hat{\tau}^* \approx \hat{\tau}_k := \left(-\frac{\alpha\lambda_N^{1/2}}{8k}\ln\left(\frac{\lambda_N}{c}\right) + \sqrt{\left(\frac{\alpha\lambda_N^{1/2}}{8k}\ln\left(\frac{\lambda_N}{c}\right)\right)^2 + (c\,\lambda_N)^{1/2}}\right)^2. \tag{32}$$

Proof From (31) we have

$$c^{-\alpha}\left[\frac{\tau^{1/2} - c^{1/2}}{\tau^{1/2} + c^{1/2}}\right]^{2k} = \lambda_N^{-\alpha}\left[\frac{\lambda_N^{1/2} - \tau^{1/2}}{\lambda_N^{1/2} + \tau^{1/2}}\right]^{2k}. \tag{33}$$

Setting $x = (c/\tau)^{1/2} < 1$ and $y = (\tau/\lambda_N)^{1/2} < 1$ by (33) we obtain

$$\left(\frac{1-x}{1+x}\right) = \left(\frac{\lambda_N}{c}\right)^{-\frac{\alpha}{2k}}\left(\frac{1-y}{1+y}\right).$$

Using (27) we solve

$$e^{-2x} = \left(\frac{\lambda_N}{c}\right)^{-\frac{\alpha}{2k}} e^{-2y}.$$

Therefore

$$-2x = -\frac{\alpha}{2k}\ln\left(\frac{\lambda_N}{c}\right) - 2y$$

which implies

$$x - y = \frac{\alpha}{4k} \ln\left(\frac{\lambda_N}{c}\right).$$

Substituting x by $(c/\tau)^{1/2}$ and y by $(\tau/\lambda_N)^{1/2}$ after some algebra we obtain

$$\tau + \frac{\alpha}{4k}\lambda_N^{1/2} \ln\left(\frac{\lambda_N}{c}\right)\tau^{1/2} - (c\,\lambda_N)^{1/2} = 0.$$

Then, solving this equation and taking the positive solution, we obtain the expression of $\hat{\tau}_k$. □

Observe that by (32), for $k \to +\infty$ we have

$$\begin{aligned}\left(\frac{\hat{\tau}_k}{\lambda_N}\right)^{1/2} &= -\frac{\alpha}{8k}\ln\left(\frac{\lambda_N}{c}\right) + \sqrt{\left(\frac{\alpha}{8k}\ln\left(\frac{\lambda_N}{c}\right)\right)^2 + \left(\frac{c}{\lambda_N}\right)^{1/2}} \\ &= -\frac{\alpha}{8k}\ln\left(\frac{\lambda_N}{c}\right) + \left(\frac{c}{\lambda_N}\right)^{1/4} + \mathcal{O}\left(\frac{1}{k^2}\right), \qquad (34)\end{aligned}$$

and therefore $\hat{\tau}_k \to \tilde{\tau}$, the asymptotically optimal parameter defined by (7). Finally, using (27) and the above expression we obtain

$$\begin{aligned}\varphi_3\left(\hat{\tau}_k\right) &= \lambda_N^{-\alpha}\left[\frac{\lambda_N^{1/2} - \hat{\tau}_k^{1/2}}{\lambda_N^{1/2} + \hat{\tau}_k^{1/2}}\right]^{2k} \\ &\leq \lambda_N^{-\alpha}\exp\left(-4k\left(\frac{\hat{\tau}_k}{\lambda_N}\right)^{1/2}\right) \\ &= \lambda_N^{-\alpha}\exp\left(-4k\left[\left(\frac{c}{\lambda_N}\right)^{1/4} - \frac{\alpha}{8k}\ln\left(\frac{\lambda_N}{c}\right)\right]\right)\left(1 + \mathcal{O}\left(\frac{1}{k}\right)\right) \\ &= \lambda_N^{-\alpha}\exp\left(-4k\left(\frac{c}{\lambda_N}\right)^{1/4}\right)\exp\left(\frac{\alpha}{2}\ln\left(\frac{\lambda_N}{c}\right)\right)\left(1 + \mathcal{O}\left(\frac{1}{k}\right)\right) \\ &= (c\,\lambda_N)^{-\alpha/2}\exp\left(-4k\left(\frac{c}{\lambda_N}\right)^{1/4}\right)\left(1 + \mathcal{O}\left(\frac{1}{k}\right)\right). \qquad (35)\end{aligned}$$

The above analysis yields the following result.

Theorem 3 *Let $\overline{k}$ be such that for each $k \geq \overline{k}$ we have $\overline{\lambda} = \overline{\lambda}(k) > \lambda_N$. Then for each $k \geq \overline{k}$, taking in* (2) *$\tau = \hat{\tau}_k$, where $\hat{\tau}_k$ is given in* (32)*, the following bound holds*

$$\left\| \mathcal{L}_N^{-\alpha} - \hat{\tau}_k^{-\alpha} R_{k-1,k}\left(\frac{\mathcal{L}_N}{\hat{\tau}_k}\right)\right\|_2 \leq 2\sin(\alpha\pi)\,(c\,\lambda_N)^{-\alpha/2} \times \exp\left(-4k\left(\frac{c}{\lambda_N}\right)^{1/4}\right)\left(1+\mathcal{O}\left(\frac{1}{k}\right)\right). \quad (36)$$

It is important to remark that, qualitatively, we have obtained the same result of [2] and reported in (3) following a completely different approach. Nevertheless the analysis presented here is quantitatively more accurate as it also provides the constant that multiplies the exponential factor. Observe moreover that the analysis of this section may be particularly useful when, in practical situation, one is forced to keep the discretization quite coarse (so that $\overline{k}$ may be rather small) and also to keep small the number of quadrature nodes k. In this case, defining $\hat{\tau}_k$ as in (32) may provide results much better than the one attainable with the asymptotically optimal choice $\tilde{\tau} = \sqrt{\lambda_{\min}(\mathcal{L}_N)\lambda_{\max}(\mathcal{L}_N)}$.

In order to compute a fairly accurate estimate of $\overline{k}$ we solve the equation $\bar{\lambda} = \lambda_N$, where $\bar{\lambda}$ is defined in (19). Neglecting the factor s_k^2 in (19) and taking $\tau = \tau_k$ as in (24), we obtain the equation

$$W\left(\frac{4k^2 e}{\alpha^2}\right) = \frac{1}{2}\ln\left(\frac{\lambda_N}{c}e^2\right).$$

Since $W(z_1) = z_2$ if and only if $z_1 = z_2 e^{z_2}$, we clearly have

$$\frac{4k^2}{\alpha^2} = \frac{1}{2}\ln\left(\frac{\lambda_N}{c}e^2\right)\left(\frac{\lambda_N}{c}\right)^{1/2}$$

from which the approximation to $\overline{k}$ easily follows. In practice, assuming to have a good estimate of the interval containing the spectrum of $\mathcal{L}_N$, one should use τ_k as in (24) whenever $k < \overline{k}$ and then switch to $\hat{\tau}_k$ as in (32) for $k \geq \overline{k}$. In other words, for bounded operators we consider the sequence

$$\tau_{k,N} = \begin{cases} \tau_k & \text{if } k < \overline{k}, \\ \hat{\tau}_k & \text{if } k \geq \overline{k}, \end{cases} \quad (37)$$

with

$$\bar{k} = \frac{\alpha}{2\sqrt{2}}\left(\ln\left(\frac{\lambda_N}{c}e^2\right)\right)^{1/2}\left(\frac{\lambda_N}{c}\right)^{1/4}. \quad (38)$$

5 Numerical experiments

In this section we present the numerical results obtained by considering two simple cases of self-adjoint positive operators. In particular, in the first example we try to

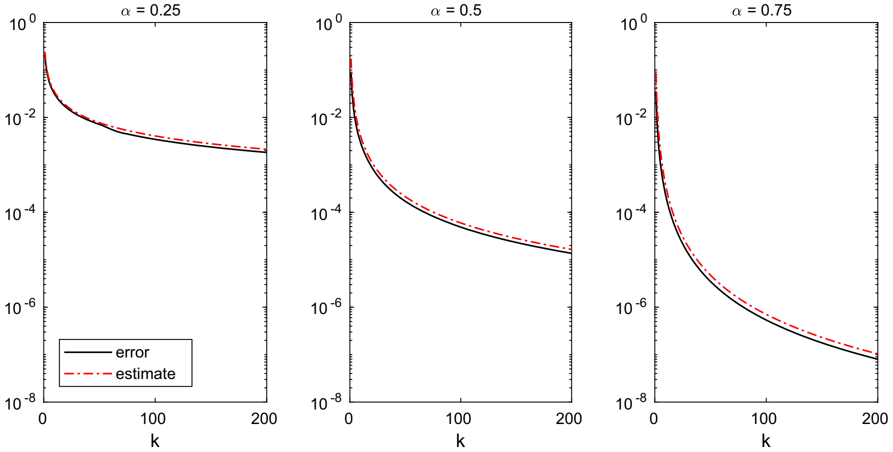

Fig. 2 Error and error bound (29) for Example 1 with $N = 100$, $p = 4$

simulate the behavior of an unbounded operator by working with a diagonal matrix with a wide spectrum. In the second one we consider the standard central difference discretization of the one dimensional Laplace operator with Dirichlet boundary conditions.

We remark that in all the experiments the weights and nodes of the Gauss–Jacobi quadrature rule are computed by using the Matlab function `jacpts` implemented in Chebfun by Hale and Townsend [13]. In addition, the errors are always plotted with respect to the Euclidean norm.

Example 1 We define $A = \mathrm{diag}(1, 2, \ldots, N)$ and $\mathcal{L}_N = A^p$ so that $\sigma(\mathcal{L}_N) \subseteq [1, N^p]$. Taking $N = 100$ and $p = 4$, in Fig. 2, for $\alpha = 0.25, 0.5, 0.75$ the error and the error bound (29) are plotted versus k, the number of points of the Gauss–Jacobi rule. It is worth noting that (29) provides excellent estimates even for small values of k, although the analysis has been made assuming that k is large (cf. Remark 1).

In Fig. 3, for $\alpha = 0.5$ we plot the error obtained using τ_k taken as in (24) and $\tilde{\tau}$ as in (7), changing the amplitude of the spectrum, that is, the value of p. In particular, we fix again $N = 100$ and take $p = 2, 3, 4$. Since the value of τ_k does not depend on the amplitude of the spectrum, there is only one curve for this value.

The figure clearly shows the improvement attainable with τ_k for k small, and moreover the deterioration of the method for very large spectra when using $\tilde{\tau}$.

Example 2 We consider the linear operator $\mathcal{L}u = -u''$, $u : [0, b] \to \mathbb{R}$, with Dirichlet boundary conditions $u(0) = u(b) = 0$. It is known that $\mathcal{L}$ has a point spectrum consisting entirely of eigenvalues

$$\mu_s = \frac{\pi^2 s^2}{b^2}, \qquad \text{for } s = 1, 2, 3, \ldots.$$

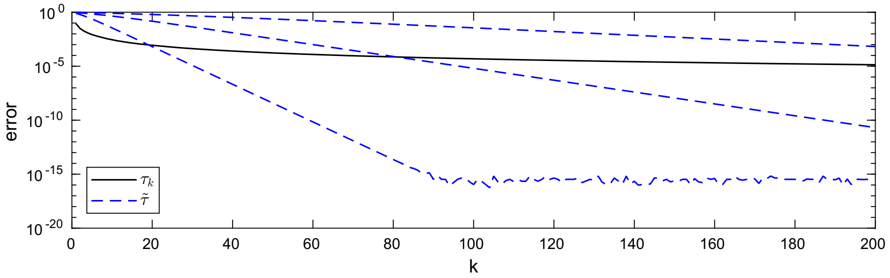

Fig. 3 Error comparison for Example 1 using $\tilde{\tau}$ as in (7) and τ_k as in (24), $p = 2, 3, 4$ (lowest to highest curve), $N = 100$ and $\alpha = 0.5$

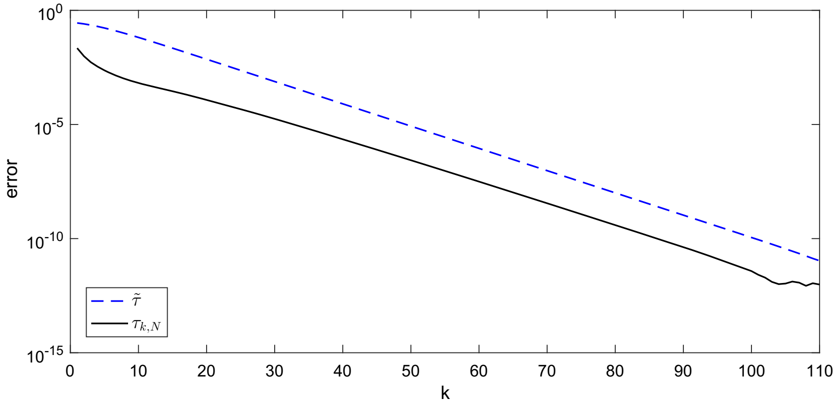

Fig. 4 Error comparison for Example 2 using $\tilde{\tau}$ as in (7) and $\tau_{k,N}$ as in (37), $N = 500$ and $\alpha = 0.5$

Using the standard central difference scheme on a uniform grid and setting $b = 1$, in this example we work with the operator

$$\mathcal{L}_N := (N+1)^2 \text{tridiag}(-1, 2, -1) \in \mathbb{R}^{N \times N}.$$

The eigenvalues are

$$\lambda_j = 4(N+1)^2 \sin^2\left(\frac{j\pi}{2(N+1)}\right), \qquad j = 1, 2, \ldots, N,$$

so that $\sigma(\mathcal{L}_N) \subseteq [\pi^2, 4(N+1)^2]$.

The aim of this example is to show the improvement that can be obtained by using the k-dependent parameter $\tau_{k,N}$ as in (37) with respect to the asymptotically optimal one $\tilde{\tau}$. By choosing $N = 500$, so that $\lambda_N \approx 10^6$, and $\alpha = 0.5$, we get $\bar{k} = 12$ (11.6 from the exact computation by (38)). In Fig. 4 the errors are reported. In Fig. 5 we also plot the values of the sequence $\tau_{k,N}$. We remark that for other choice of α the results are qualitatively identical.

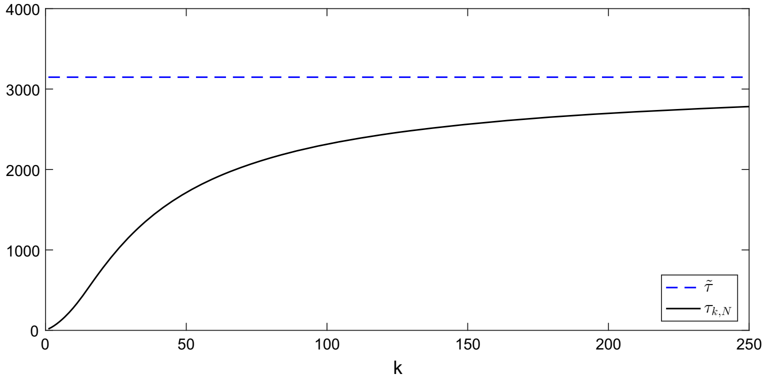

Fig. 5 Selected values for $\tau_{k,N}$ defined by (37) for Example 2 with $N = 500$ and $\alpha = 0.5$

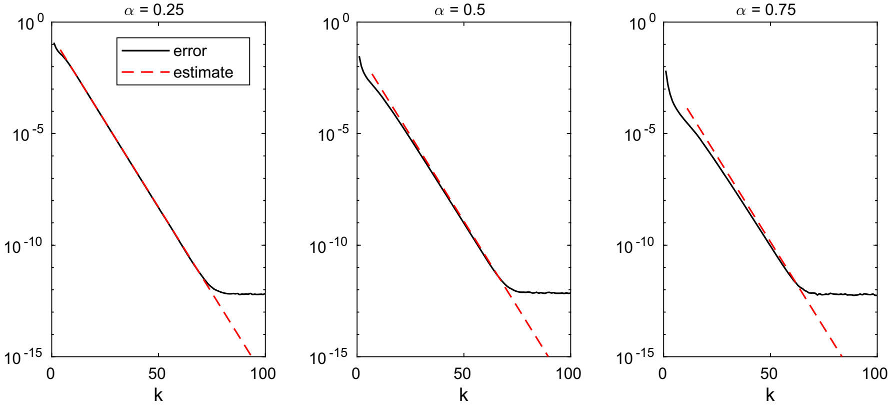

Fig. 6 Error and error estimate (36) for Example 2 with $N = 200$

Finally, still working with this example, we show the accuracy of the bound (36) for $\alpha = 0.25, 0.5, 0.75$. The results are reported in Fig. 6.

6 Conclusions

In this paper we have considered rational approximations of fractional powers of unbounded positive operators obtained by exploiting the connection between Gauss–Jacobi quadrature on Markov functions and Padé approximants. Using classical results in approximation theory, we have provided very sharp a priori estimates of the truncation errors that allow to properly define the parameter τ. The numerical experiments confirm that such analysis improves some existing results.

On the other hand, in the paper we have not considered the computational issues behind this kind of approximations, since they are strictly dependent on the operator $\mathcal{L}$ or its discretization. Clearly one inversion (or one linear system if one needs to approximate $\mathcal{L}^{-\alpha} f$) is necessary at each step so that a suitable preconditioning approach should be employed. In this way the stagnation around 10^{-12} observed in Fig. 6 could be overtaken.

References

1. Aceto, L., Magherini, C., Novati, P.: On the construction and properties of m-step methods for FDEs. SIAM J. Sci. Comput. **37**, A653–A675 (2015)
2. Aceto, L., Novati, P.: Rational approximation to the fractional Laplacian operator in reaction-diffusion problems. SIAM J. Sci. Comput. **39**, A214–A228 (2017)
3. Aceto, L., Novati, P.: Efficient implementation of rational approximations to fractional differential operators. J. Sci. Comput. **76**, 651–671 (2018)
4. Bhatia, R.: Matrix Analysis, vol. 169 of Graduate Texts in Mathematics. Springer, New York (1997)
5. Bonito, A., Pasciak, J.E.: Numerical approximation of fractional powers of elliptic operators. Math. Comp. **84**, 2083–2110 (2015)
6. Clark, M.A., Kennedy, A.D.: Accelerating dynamical-fermion computations using the rational hybrid Monte Carlo algorithm with multiple pseudofermion fields. Phys. Rev. Lett. **98**(1), 051601 (2007)
7. Elliot, D.: Truncation errors in Padé approximations to certain functions: an alternative approach. Math. Comput. **21**, 398–406 (1967)
8. Iannazzo, B.: A family of rational iterations and its application to the computation of the matrix pth root. SIAM J. Matrix Anal. Appl. **30**, 1445–1462 (2008)
9. Iannazzo, B.: On the Newton method for the matrix pth root. SIAM J. Matrix Anal. Appl. **28**, 503–523 (2006)
10. Erdélyi, A., Magnus, W., Oberhettinger, F., Tricomi, F.G.: Higher Transcendental Functions, vol. 1. McGraw-Hill, New York (1953)
11. Fasi, M., Iannazzo, B.: Computing the weighted geometric mean of two large-scale matrices and its inverse times a vector. SIAM J. Matrix Anal. Appl. **39**, 178–203 (2018)
12. Frommer, A., Güttel, S., Schweitzer, M.: Efficient and stable Arnoldi restarts for matrix functions based on quadrature. SIAM J. Matrix Anal. Appl. **35**, 661–683 (2014)
13. Hale, N., Townsend, A.: Fast and accurate computation of Gauss-Legendre and Gauss-Jacobi quadrature nodes and weights. SIAM J. Sci. Comput. **35**, A652–A672 (2013)
14. Harizanov, S., Lazarov, R., Margenov, S., Marinov, P., Vutov, Y.: Optimal solvers for linear systems with fractional powers of sparse SPD matrices. Numerical Linear Algebra Appl. **25**, 115–128 (2018)
15. Harizanov, S., Lazarov, R., Marinov, P., Margenov, S., Pasciak, J.: Comparison analysis on two numerical methods for fractional diffusion problems based on rational approximations of t^γ, $0 \le t \le 1$. arXiv:1805.00711v1 (2018)
16. Ilić, M., Liu, F., Turner, I., Anh, V.: Numerical approximation of a fractional-in-space diffusion equation I. Fract. Calc. Appl. Anal. **8**, 323–341 (2005)
17. Kwaśnicki, M.: Ten equivalent definitions of the fractional Laplace operator. Fract. Calc. Appl. Anal. **20**, 7–51 (2017)
18. Lebedev, N.N.: Special Functions and Their Applications. Prentice-Hall Inc., Englewood Cliffs (1965)
19. Moret, I., Novati, P.: Krylov subspace methods for functions of fractional differential operators. Math. Comput. **88**, 293–312 (2019)
20. Novati, P.: Numerical approximation to the fractional derivative operator. Numer. Math. **127**, 539–566 (2014)
21. Sadeghi, A., Ismail, A.I.M., Ahmad, A.: Computing the pth roots of a matrix with repeated eigenvalues. Appl. Math. Sci. **5**, 2645–2661 (2011)
22. Ringrose, J.R.: Compact Non-Self-Adjoint Operators. Van Nostrand Reinhold Company, London (1971)
23. Tatsuoka, F., Sogabe, T., Miyatake, Y., Zhang, S.: A cost-efficient variant of the incremental Newton iteration for the matrix pth root. J. Math. Res. Appl. **37**, 97–106 (2017)