Research Article

Eleonora Denich* and Paolo Novati

A Gaussian Method for the Square Root of Accretive Operators

https://doi.org/10.1515/cmam-2022-0033
Received February 3, 2022; revised May 9, 2022; accepted August 17, 2022

Abstract: We consider the approximation of the inverse square root of regularly accretive operators in Hilbert spaces. The approximation is of rational type and comes from the use of the Gauss–Legendre rule applied to a special integral formulation of the fractional power. We derive sharp error estimates, based on the use of the numerical range, and provide some numerical experiments. For practical purposes, the finite-dimensional case is also considered. In this setting, the convergence is shown to be of exponential type. The method is also tested for the computation of a generic fractional power.

Keywords: Fractional Powers, Regularly Accretive Operators, Gaussian Quadrature Rule

MSC 2010: 47A58, 65F60, 65D32

1 Introduction

Let $\mathcal{H}$ be a generic Hilbert space with scalar product denoted by $\langle\cdot,\cdot\rangle$ and corresponding norm $\|u\| = \langle u,u\rangle^{\frac{1}{2}}$, $u \in \mathcal{H}$. Given a linear operator $\mathcal{L}$ acting in $\mathcal{H}$, in this work, we are interested in the numerical approximation of $\mathcal{L}^{-\frac{1}{2}}$, where $\mathcal{L}$ is assumed to be regularly accretive, that is, associated with a regular sesquilinear form (see [17] for a background). It is known that such operators are unbounded and satisfy

$$|\mathfrak{I}\langle \mathcal{L}u,u\rangle| \le \eta\, \mathfrak{R}\langle \mathcal{L}u,u\rangle, \tag{1.1}$$

for some $\eta \ge 0$, where the symbols $\mathfrak{I}$ and $\mathfrak{R}$ indicate the imaginary and the real part respectively. For $a \ge 0$ and $\beta < \frac{1}{2}$, let

$$\Sigma_{\beta,a} = \{z \in \mathbb{C} : |\arg(z-a)| \le \beta\pi\} \tag{1.2}$$

be the sector symmetric with respect to the real axis with vertex in a and semiangle $\beta\pi$. Denoting by $F(\mathcal{L})$ the numerical range of $\mathcal{L}$, that is,

$$F(\mathcal{L}) = \left\{ z \in \mathbb{C} : z = \frac{\langle u,\mathcal{L}u\rangle}{\langle u,u\rangle},\ u \in \mathcal{H},\ u \ne 0 \right\},$$

referring to (1.1), it is also known that (see [17, Theorem 2.2])

$$F(\mathcal{L}) \subseteq \Sigma_{\frac{\arctan(\eta)}{\pi},0}.$$

In this setting, for $\alpha \in (0,1)$, the fractional power is defined by (see [5])

$$\mathcal{L}^{-\alpha}v = \frac{\sin(\alpha\pi)}{\pi}\int_0^\infty \varsigma^{-\alpha}(\varsigma\mathcal{I}+\mathcal{L})^{-1}v\,d\varsigma, \quad v \in \mathcal{H}, \tag{1.3}$$

where $\mathcal{I}$ is the identity operator in $\mathcal{H}$.

***Corresponding author: Eleonora Denich,** Dipartimento di Matematica e Geoscienze, Università di Trieste, Trieste, Italy, e-mail: eleonora.denich@phd.units.it. https://orcid.org/0000-0003-2403-753X
Paolo Novati, Dipartimento di Matematica e Geoscienze, Università di Trieste, Trieste, Italy, e-mail: novati@units.it

Starting from this representation, in this work, we consider some changes of variable that, in the scalar case, lead to the formula

$$\lambda^{-\alpha} = \frac{\sin(\alpha\pi)}{\pi}\tau^{1-\alpha}\left(\frac{2^{\frac{1-\alpha}{\alpha}}}{\alpha}\int_{-1}^{1}\frac{1}{2^{\frac{1}{\alpha}}\tau+\lambda(t+1)^{\frac{1}{\alpha}}}\,dt+\frac{2^{\frac{\alpha}{1-\alpha}}}{1-\alpha}\int_{-1}^{1}\frac{1}{\tau(t+1)^{\frac{1}{1-\alpha}}+2^{\frac{1}{1-\alpha}}\lambda}\,dt\right), \tag{1.4}$$

where $\tau > 0$ is a parameter that allows to handle the case in which λ is replaced by an operator $\mathcal{L}$ such that $F(\mathcal{L}) \subset \Sigma_{\beta,a}$ (Section 4). For the approximation of (1.4), we employ the Gauss–Legendre rule. Working with the n-point formula for both integrals, we implicitly construct a rational form of type

$$\mathcal{R}_{2n-1,2n}(\lambda) = \frac{p_{2n-1}(\lambda)}{q_{2n}(\lambda)}, \quad p_{2n-1} \in \Pi_{2n-1},\ q_{2n} \in \Pi_{2n},$$

such that $\mathcal{R}_{2n-1,2n}(\lambda) \cong \lambda^{-\alpha}$. Here and below, we use the symbol $\cong$ to indicate a generic approximation.

Similar approaches, based on quadrature rules arising from the Dunford–Taylor integral representation of $\lambda^{-\alpha}$, have been considered for instance in [1–4, 7, 8, 21]. Other methods that rely on the best uniform rational approximations of functions closely related to $\lambda^{-\alpha}$ have been treated in [11–14]. Methods based on the parabolic reformulation of fractional diffusion equations have been analyzed in [19, 20]. As pointed out in [15], they can still be interpreted as rational approximations of $\lambda^{-\alpha}$. We also quote here [10] for a very recent survey. We remark that, except for [7], in all these papers, the basic assumption has been to work with a self-adjoint operator.

As for the method considered in this paper, in order to properly define the parameter τ appearing in (1.4), we need to derive accurate error estimates. Unfortunately, this is only possible whenever the integrand functions are both analytic; indeed, for non-analytic functions, the existing error formulas only describe the qualitative behavior of the error decay (see e.g. [6], [18, Section 4]). For this reason, in this paper, we study the error and the proper definition of τ only for $\alpha = \frac{1}{2}$, which ensures that both the integrand functions in (1.4) are analytic. The analysis developed for the derivation of τ allows to show that, in the operator norm, the convergence rate is of type $\frac{(\ln n)^2}{n^4}$, where n is the number of quadrature points. We do not claim that the proposed method is the fastest since, by using an exponential transform in (1.3) and then the trapezoidal rule, it is possible to achieve an exponential decay for the error, as shown in [7]. Nevertheless, the method has some potentials. Indeed, the rate of convergence is independent of the angle of the sector containing$F(\mathcal{L})$ (it only affects the error constant), the initial convergence is very fast because of the factor $\frac{1}{n^4}$, and, in addition, we have been able to derive a sharp error estimate that allows to select a priori the number of quadrature points necessary to achieve a certain accuracy.

For practical purposes, still for $\alpha = \frac{1}{2}$, we have also analyzed the behavior of the method in finite dimension, that is, in the case of bounded sectorial operators $\mathcal{L}_N$, showing that the decay of the error is of type

$$\frac{c_1}{\|\mathcal{L}_N\|^{\frac{1}{4}}}\exp\left(-c_2\frac{n}{\|\mathcal{L}_N\|^{\frac{1}{8}}}\right),$$

where c_1 and c_2 are constants depending on the angle of the sector containing $F(\mathcal{L}_N)$, and $\|\cdot\|$ is the spectral norm. It is interesting to observe that, in the case of $\mathcal{L}_N$ symmetric and positive definite, the above formula can be rewritten replacing $\|\mathcal{L}_N\|$ with the spectral condition number $\kappa(\mathcal{L}_N)$, resulting in a clear improvement with respect for instance to the Gauss–Jacobi approach [2], where the above formula still holds, but with $\|\mathcal{L}_N\|$ replaced by $\|\mathcal{L}_N\|^2$.

Without a rigorous error analysis, we have also considered the use of the Gauss–Legendre method for the computation of $\mathcal{L}^{-\alpha}$, with $\alpha \neq \frac{1}{2}$. The method appears to work rather fine even if, in this case, the choice of the parameter τ (see (1.4)), which plays a crucial role for the convergence behavior, is empirical.

The paper is organized as follows. In Section 2, we derive the integral representation (1.4) and approximate it by using the Gauss–Legendre rule. In Section 3, we develop the error analysis in the scalar case by studying the poles of the integrand functions in the complex plane. In Section 4, we generalize the analysis for regularly accretive operators, by considering the error behavior on the boundaries of the sector containing the numerical range of the operator. In Section 5, working with bounded operators, we improve the error estimates previously obtained. Finally, in Section 6, we present some experiments for the general case of $\alpha \in (0, 1)$.

2 The Gauss–Legendre Approach

Starting from the integral representation (1.3), we consider the change of variable $\varsigma = x^{-2}$ (see [8]) that leads to

$$\lambda^{-\alpha} = \frac{2\sin(\alpha\pi)}{\pi}\int_0^{\infty} \frac{x^{2\alpha-1}}{1+x^2\lambda}\,dx, \quad \lambda \in \mathbb{C}\setminus(-\infty,0].$$

Then we split the above integral as follows:

$$\lambda^{-\alpha} = \frac{2\sin(\alpha\pi)}{\pi}\left(\int_{\frac{1}{\sqrt{\tau}}}^{\infty} \frac{x^{2\alpha-1}}{1+x^2\lambda}\,dx + \int_0^{\frac{1}{\sqrt{\tau}}} \frac{x^{2\alpha-1}}{1+x^2\lambda}\,dx\right),$$

where $\tau \geq 1$ is a certain parameter whose meaning will be explained later. By using the changes of variable

$$x = \frac{1}{\sqrt{\tau}} y^{-\frac{1}{2(1-\alpha)}} \quad \text{and} \quad x = \frac{1}{\sqrt{\tau}} y^{\frac{1}{2\alpha}}$$

for the first and the second integral respectively, we have that

$$\lambda^{-\alpha} = \frac{\sin(\alpha\pi)}{\pi}\tau^{1-\alpha}\left(\frac{1}{\alpha}\int_0^1 \frac{1}{\tau+\lambda y^{\frac{1}{\alpha}}}\,dy + \frac{1}{1-\alpha}\int_0^1 \frac{1}{\tau y^{\frac{1}{1-\alpha}}+\lambda}\,dy\right). \tag{2.1}$$

Finally, for both integrals in (2.1), we apply the change of variable

$$y = \frac{t+1}{2}$$

and obtain the integral representation (1.4), that is,

$$\lambda^{-\alpha} = \frac{\sin(\alpha\pi)}{\pi}\tau^{1-\alpha}\left(\frac{2^{\frac{1-\alpha}{\alpha}}}{\alpha} I^{(1)}(\lambda) + \frac{2^{\frac{\alpha}{1-\alpha}}}{1-\alpha} I^{(2)}(\lambda)\right), \tag{2.2}$$

where

$$I^{(1)}(\lambda) := \int_{-1}^{1} \frac{1}{2^{\frac{1}{\alpha}}\tau + \lambda(t+1)^{\frac{1}{\alpha}}}\,dt, \quad I^{(2)}(\lambda) := \int_{-1}^{1} \frac{1}{\tau(t+1)^{\frac{1}{1-\alpha}} + 2^{\frac{1}{1-\alpha}}\lambda}\,dt. \tag{2.3}$$

We remark that representation (2.1) has also been used in [21] as starting point for the construction of other methods. Using the n-point Gauss–Legendre rule, formula (2.2) is approximated as

$$\lambda^{-\alpha} \cong \frac{\sin(\alpha\pi)}{\pi}\tau^{1-\alpha}\left(\frac{2^{\frac{1-\alpha}{\alpha}}}{\alpha} I_n^{(1)}(\lambda) + \frac{2^{\frac{\alpha}{1-\alpha}}}{1-\alpha} I_n^{(2)}(\lambda)\right), \tag{2.4}$$

where

$$I_n^{(1)}(\lambda) := \sum_{j=1}^{n} \omega_j\left(2^{\frac{1}{\alpha}}\tau + \lambda(t_j+1)^{\frac{1}{\alpha}}\right)^{-1}, \quad I_n^{(2)}(\lambda) := \sum_{j=1}^{n} \omega_j\left(\tau(t_j+1)^{\frac{1}{1-\alpha}} + 2^{\frac{1}{1-\alpha}}\lambda\right)^{-1}, \tag{2.5}$$

in which t_j, ω_j, $j = 1,\ldots,n$, are respectively the nodes and the weights of the Gaussian rule. As mentioned in the introduction, we observe that (2.4) is in fact a rational approximation $\mathcal{R}_{2n-1,2n}(\lambda)$ of $\lambda^{-\alpha}$, where

$$\mathcal{R}_{2n-1,2n}(\lambda) = \frac{p_{2n-1}(\lambda)}{q_{2n}(\lambda)}, \quad p_{2n-1} \in \Pi_{2n-1},\ q_{2n} \in \Pi_{2n}.$$

3 General Error Analysis

As mentioned in the introduction, the error analysis is only given for the case $\alpha = \frac{1}{2}$. Let us consider the transform

$$\psi(w) = \frac{1}{2}\left(w + \frac{1}{w}\right), \quad w \in \mathbb{C},\ |w| > 1,$$

that conformally maps the exterior of the unit circle onto the exterior of the interval $[-1, 1]$. This map is usually called Joukowsky transform. The image of the circle $|w| = s$, that is,

$$\Psi_s = \{z \in \mathbb{C} : z = \psi(se^{i\theta}),\ \theta \in [0, 2\pi]\},$$

is an ellipse of the complex plane with foci in ± 1. We denote by $\mathcal{E} = \{\Psi_s : s > 1\}$ the family of all these ellipses.

Let f be a generic function analytic in an open set containing $[-1, 1]$. Let moreover

$$I(f) = \int_{-1}^{1} f(t)\, dt,$$

and let $I_n(f)$ be its n-point Gauss–Legendre approximation. Following the analysis given in [6], assume that $\Psi_s \in \mathcal{E}$ is such that f is analytic in the interior of Ψ_s, except for a pair of simple poles t_0 and its conjugate $\bar{t}_0$. Then

$$I(f) - I_n(f) \cong -4\pi\, \Re\{r(t_0 + \sqrt{t_0^2 - 1})^{-2n}\}, \tag{3.1}$$

where r is the residue of f at t_0 and the root has to be chosen such that $S := |t_0 + \sqrt{t_0^2 - 1}| > 1$. It is important to observe that S is just the radius of the circle centered at 0 that, through the Joukowsky transform, is mapped onto the ellipse $\Psi_S \in \mathcal{E}$ passing through t_0 and $\bar{t}_0$. Clearly, by (3.1), the rate of convergence grows with S that, roughly speaking, handles the distance of the poles from the interval $[-1, 1]$. Having at disposal the above general result, we can estimate the error of approximation (2.4) by studying separately the poles of our integrand functions (cf. (2.3))

$$f^{(1)}(t) = \frac{1}{4\tau + \lambda(t+1)^2} \quad \text{and} \quad f^{(2)}(t) = \frac{1}{\tau(t+1)^2 + 4\lambda}. \tag{3.2}$$

As for the function $f^{(1)}$, it is easy to see that the poles are given by

$$t_{0,1}^{(1)} = \pm 2\left(\frac{\tau}{\lambda}\right)^{\frac{1}{2}} i - 1 \tag{3.3}$$

so that $t_0^{(1)} = \overline{t_1^{(1)}}$. Similarly, for the function $f^{(2)}$, we have

$$t_{0,1}^{(2)} = \pm 2\left(\frac{\lambda}{\tau}\right)^{\frac{1}{2}} i - 1, \tag{3.4}$$

and therefore $t_0^{(2)} = \overline{t_1^{(2)}}$. By using (3.1), and defining $e_n^{(i)}(\lambda) = I^{(i)}(\lambda) - I_n^{(i)}(\lambda)$, $i = 1, 2$ (cf. (2.2) and (2.4)), we obtain

$$e_n^{(i)}(\lambda) \cong -4\pi \Re\{r^{(i)}(t_0^{(i)} + \sqrt{(t_0^{(i)})^2 - 1})^{-2n}\}.$$

Since

$$r^{(i)} = \lim_{t \to t_0^{(i)}} (t - t_0^{(i)}) f^{(i)}(t), \quad i = 1, 2,$$

by (3.2), (3.3), (3.4), we find

$$r^{(1)} = \frac{1}{4\sqrt{\tau}} \lambda^{-\frac{1}{2}} i \quad \text{and} \quad r^{(2)} = -\frac{1}{4\sqrt{\tau}} \lambda^{-\frac{1}{2}} i,$$

and therefore

$$|e_n^{(i)}(\lambda)| \cong 4\pi |r^{(i)}| {S^{(i)}}^{-2n} = \frac{\pi}{\sqrt{\tau}} |\lambda|^{-\frac{1}{2}} {S^{(i)}}^{-2n} =: \Phi^{(i)}(\tau, \lambda), \tag{3.5}$$

where

$$S^{(i)} = \left|t_0^{(i)} + \sqrt{(t_0^{(i)})^2 - 1}\right|, \quad i = 1, 2. \tag{3.6}$$

As for the total error

$$E_n(\lambda) = \frac{4\sqrt{\tau}}{\pi} (e_n^{(1)}(\lambda) + e_n^{(2)}(\lambda)) \tag{3.7}$$

(cf. (2.2) and (2.4)), we then have the estimate

$$|E_n(\lambda)| \cong 4|\lambda|^{-\frac{1}{2}} ({S^{(1)}}^{-2n} + {S^{(2)}}^{-2n}). \tag{3.8}$$

In Figure 1, we show the accuracy of the above formula for some values of λ and $\tau = 2$.

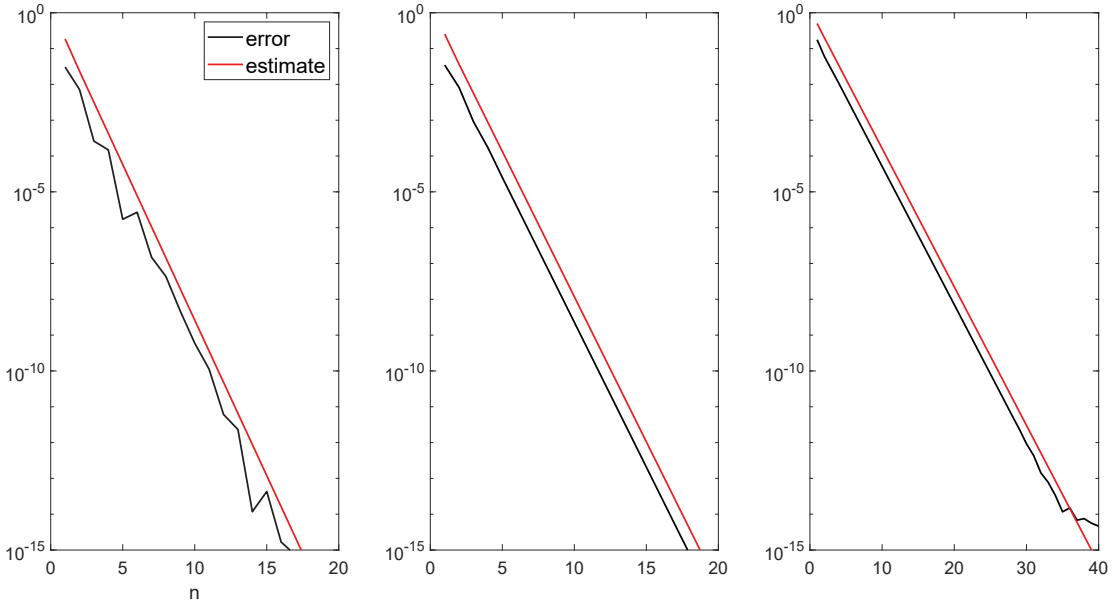

Figure 1: Error and error estimate (3.8) for $\lambda = 10$, $\lambda = 5 + 5i$, $\lambda = -5 + 10i$, from left to right, and $\tau = 2$.

4 Estimates for Operators

Let $\mathcal{H}$ be a generic Hilbert space, and let $\mathcal{L} : \mathcal{H} \to \mathcal{H}$ be a regularly accretive operator such that $F(\mathcal{L}) \subseteq \Sigma_{\beta,1}$ (see (1.2)). It is known that, for any function f analytic in $F(\mathcal{L})$, it holds

$$\|f(\mathcal{L})\|_{\mathcal{H}\to\mathcal{H}} \le K \max_{\lambda\in F(\mathcal{L})} |f(\lambda)|, \tag{4.1}$$

where $2 \le K \le 1 + \sqrt{2}$ is the absolute constant studied in [9]. We remark that if $\mathcal{L}$ is self-adjoint, then $\beta = 0$ and $K = 1$. As a consequence, since the poles of the approximation $\mathcal{R}_{2n-1,2n}(\lambda)$ are all in $\mathbb{R}^-$ (cf. (2.4) and (2.5)), we can consider the bound

$$\|E_n(\mathcal{L})\|_{\mathcal{H}\to\mathcal{H}} = \|\mathcal{L}^{-\alpha} - \mathcal{R}_{2n-1,2n}(\mathcal{L})\|_{\mathcal{H}\to\mathcal{H}} \le K \max_{\lambda\in\Sigma_{\beta,1}} |\lambda^{-\alpha} - \mathcal{R}_{2n-1,2n}(\lambda)|.$$

When studying the behavior of the method applied to $\lambda^{-\frac{1}{2}}$ for $\lambda \in \Sigma_{\beta,1}$, that is, $\lambda = 1 + \rho e^{i\theta\pi}$, $\rho \ge 0$, $|\theta| \le \beta$, it is rather evident (and will be confirmed by the following analysis) that, for a fixed ρ, moving θ from 0 to β causes a progressive slow down. Further, since $\lambda^{-\alpha} - \mathcal{R}_{2n-1,2n}(\lambda)$ is analytic in $\Sigma_{\beta,1}$, by the maximum modulus principle, it is sufficient to consider the scalar error on $\Gamma_\beta = \Gamma_\beta^+ \cup \Gamma_\beta^-$, where

$$\Gamma_\beta^+ = \{z \in \mathbb{C} : z = 1 + \rho e^{i\beta\pi},\ \rho \ge 0\} \quad \text{and} \quad \Gamma_\beta^- = \{z \in \mathbb{C} : z = 1 + \rho e^{-i\beta\pi},\ \rho \ge 0\}.$$

Therefore, by using (3.7), we have that

$$\max_{\lambda\in\Sigma_{\beta,1}} |\lambda^{-\alpha} - \mathcal{R}_{2n-1,2n}(\lambda)| = \max_{\lambda\in\Gamma_\beta} |\lambda^{-\alpha} - \mathcal{R}_{2n-1,2n}(\lambda)| \le \frac{4\sqrt{\tau}}{\pi}\Big(\max_{\lambda\in\Gamma_\beta} |e_n^{(1)}(\lambda)| + \max_{\lambda\in\Gamma_\beta} |e_n^{(2)}(\lambda)|\Big).$$

At this point, for $\lambda \in \Gamma_\beta$, we consider (cf. (3.3) and (3.4)) the maps

$$\chi^{(1)}(\lambda) := 2\left(\frac{\tau}{\lambda}\right)^{\frac{1}{2}} i - 1 \quad \text{and} \quad \chi^{(2)}(\lambda) := 2\left(\frac{\lambda}{\tau}\right)^{\frac{1}{2}} i - 1,$$

which define the boundaries of the regions of the poles (with positive imaginary part) of the functions $f^{(1)}(t)$ and $f^{(2)}(t)$ respectively. These regions are plotted in Figure 2. The symmetry of Γ_β with respect to the real axis leads to the symmetry of $\chi^{(i)}(\Gamma_\beta)$ with respect to the line $\Re(z) = -1$. Indeed, for each fixed $\rho > 0$, the two

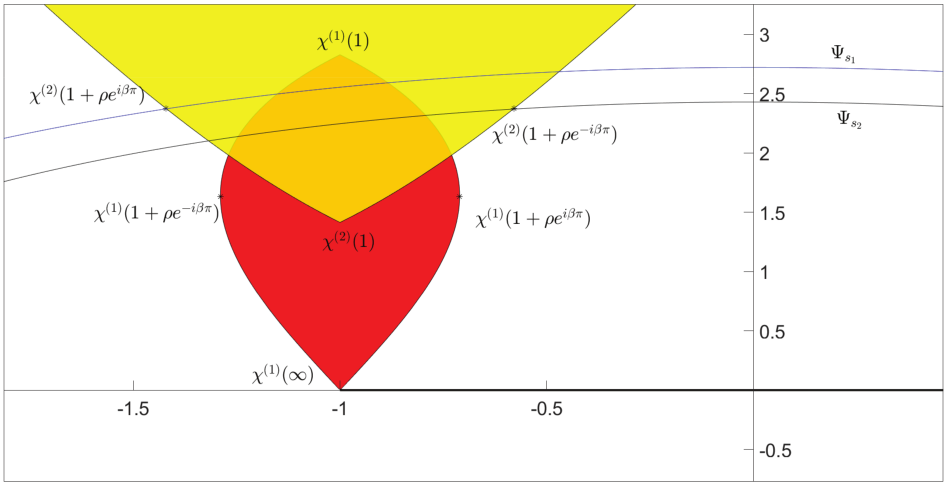

Figure 2: The regions of the poles (with positive imaginary part) of the functions $f^{(1)}(t)$ (red) and $f^{(2)}(t)$ (yellow) for $\beta = \frac{1}{6}$ and $\tau = 2$, and the ellipses Ψ_{s_1} and Ψ_{s_2}.

points $\chi^{(i)}(1+\rho e^{\pm i\beta\pi})$, $i = 1, 2$, are symmetric with respect to the line $\Re(z) = -1$. The ellipse passing through the one with the real part greater than -1 is the smallest so that, as already observed, it leads to the worst case in terms of rate of convergence (cf. (3.1)). Just for clarity, in Figure 2, we also plot the two ellipses $\Psi_{s_1}, \Psi_{s_2} \in \mathcal{E}$ passing through $\chi^{(2)}(1+\rho e^{\pm i\beta\pi})$, where

$$s_1 = \left|\chi^{(2)}(1+\rho e^{i\beta\pi}) + \sqrt{(\chi^{(2)}(1+\rho e^{i\beta\pi}))^2 - 1}\right|,$$
$$s_2 = \left|\chi^{(2)}(1+\rho e^{-i\beta\pi}) + \sqrt{(\chi^{(2)}(1+\rho e^{-i\beta\pi}))^2 - 1}\right|.$$

As a consequence, since $\Re(\chi^{(1)}(\lambda)) \geq -1$ for $\lambda \in \Gamma_\beta^+$, and $\Re(\chi^{(2)}(\lambda)) \geq -1$ for $\lambda \in \Gamma_\beta^-$, we have that

$$\max_{\lambda\in\Gamma_\beta}|e_n^{(1)}(\lambda)| = \max_{\lambda\in\Gamma_\beta^+}|e_n^{(1)}(\lambda)| \quad \text{and} \quad \max_{\lambda\in\Gamma_\beta}|e_n^{(2)}(\lambda)| = \max_{\lambda\in\Gamma_\beta^-}|e_n^{(2)}(\lambda)|.$$

Therefore, we can finally write

$$\begin{aligned}\max_{\lambda\in\Sigma_\beta}|\lambda^{-\alpha} - R_{2n-1,2n}(\lambda)| &\leq \frac{4\sqrt{\tau}}{\pi}\max_{\rho\geq 0}(|e_n^{(1)}(1+\rho e^{i\beta\pi})| + |e_n^{(2)}(1+\rho e^{-i\beta\pi})|)\\ &\cong \frac{4\sqrt{\tau}}{\pi}\max_{\rho\geq 0}(\Phi^{(1)}(\tau, 1+\rho e^{i\beta\pi}) + \Phi^{(2)}(\tau, 1+\rho e^{-i\beta\pi})),\end{aligned} \tag{4.2}$$

where we have used relation (3.5).

4.1 Error Behavior on Γ_β

Experimentally, working with operators with large spectrum, one observes that, in general, the parameter τ must be chosen quite large to achieve a good rate of convergence. For this reason, from now, we assume $\tau \gg 1$. By considering the functions

$$\Phi^{(1)}(\tau, 1+\rho e^{i\beta\pi}) \quad \text{and} \quad \Phi^{(2)}(\tau, 1+\rho e^{-i\beta\pi})$$

for $\rho \geq 0$ (cf. (4.2), (3.5)), it can be observed that, with respect to ρ, $\Phi^{(1)}(\tau, 1+\rho e^{i\beta\pi})$ initially grows, reaches a maximum at a certain $\rho^\star \gg \tau$, and then goes to 0 as $\rho \to +\infty$. On the other side, $\Phi^{(2)}(\tau, 1+\rho e^{-i\beta\pi})$ may show two kinds of behavior, depending on the angle $\beta\pi$. In particular, for $\beta \leq \beta^\star$ ($\beta^\star \cong \frac{1}{4}$), $\Phi^{(2)}(\tau, 1+\rho e^{-i\beta\pi})$ is monotone decreasing, whereas for $\beta > \beta^\star$, it initially grows, reaches a maximum at a certain $\bar{\rho} \ll \tau$, and then is monotone decreasing. We refer to Appendix A for the details concerning $\beta^\star$. Similar to the analysis

given in [2], the idea is then to define τ so that

$$\Phi^{(1)}(\tau, 1+\rho^\star e^{i\beta\pi}) = \Phi^{(2)}(\tau, 1+\bar\rho e^{-i\beta\pi}), \tag{4.3}$$

where we should set $\bar\rho = 0$ for $\beta \le \beta^\star$.

Remark 1. The reason why we impose (4.3) is that, numerically, one observes that (cf. (4.2))

$$\Phi^{(1)}(\tau, 1+\rho e^{i\beta\pi}) + \Phi^{(2)}(\tau, 1+\rho e^{-i\beta\pi}) \cong \max_{\rho\ge 0}(\Phi^{(1)}(\tau, 1+\rho e^{i\beta\pi}), \Phi^{(2)}(\tau, 1+\rho e^{-i\beta\pi})).$$

4.1.1 Approximation of $\rho^\star$

In what follows, we use the symbol ~ to relate functions asymptotically equal in the usual sense. Since $\rho^\star \gg 1$, in order to study the function $\Phi^{(1)}(\tau, 1+\rho e^{i\beta\pi})$, we first consider the approximation

$$t_0^{(1)} = 2\Big(\frac{\tau}{1+\rho e^{i\beta\pi}}\Big)^{\frac12} i - 1 \sim 2\Big(\frac{\tau}{\rho}\Big)^{\frac12} e^{i\frac{\pi}{2}(1-\beta)} - 1, \quad \rho\to\infty. \tag{4.4}$$

As a consequence, for the term $S^{(1)}$ (see (3.5) and (3.6)), we have the following result.

Proposition 1. *It holds*

$$S^{(1)} \sim 1 + \sqrt{2} C_\beta \Big(\frac{\tau}{\rho}\Big)^{\frac14}, \quad \rho\to\infty, \tag{4.5}$$

where

$$C_\beta = \sqrt{2}\cos\Big(\frac{\pi}{4}(\beta+1)\Big). \tag{4.6}$$

Proof. Assuming that $\rho \gg \tau$, by (3.6), (4.4) and using the first order approximation

$$\sqrt{1+x} = 1 + \frac{x}{2} + \mathcal{O}(x^2) \quad \text{for } x\to 0, \tag{4.7}$$

we can write

$$\begin{aligned}
S^{(1)} &\sim \left| 2\Big(\frac{\tau}{\rho}\Big)^{\frac12} e^{i\frac{\pi}{2}(1-\beta)} - 1 + \sqrt{4\Big(\frac{\tau}{\rho}\Big) e^{i\pi(1-\beta)} - 4\Big(\frac{\tau}{\rho}\Big)^{\frac12} e^{i\frac{\pi}{2}(1-\beta)}} \right| \\
&\sim \left| -1 + 2i\Big(\frac{\tau}{\rho}\Big)^{\frac14} e^{i\frac{\pi}{4}(1-\beta)} \right| \\
&= \sqrt{\Big[-1+\sqrt{2}\Big(\frac{\tau}{\rho}\Big)^{\frac14}\Big(\sin\Big(\frac{\beta}{4}\pi\Big) - \cos\Big(\frac{\beta}{4}\pi\Big)\Big)\Big]^2 + 2\Big(\frac{\tau}{\rho}\Big)^{\frac12}\Big(\sin\Big(\frac{\beta}{4}\pi\Big) + \cos\Big(\frac{\beta}{4}\pi\Big)\Big)^2} \\
&\sim \sqrt{1 + 2\sqrt{2}\Big(\frac{\tau}{\rho}\Big)^{\frac14}\Big(\cos\Big(\frac{\beta}{4}\pi\Big) - \sin\Big(\frac{\beta}{4}\pi\Big)\Big)},
\end{aligned}$$

which leads to (4.5) because

$$\cos\Big(\frac{\beta}{4}\pi\Big) - \sin\Big(\frac{\beta}{4}\pi\Big) = \sqrt{2}\cos\Big(\frac{\pi}{4}(\beta+1)\Big). \qquad \square$$

Note that $0.54 \cong \sqrt{1-\frac{\sqrt{2}}{2}} < C_\beta \le 1$ for $0 \le \beta < \frac12$ ($C_\beta = 1$ for $\beta = 0$). At this point, we look for the local maximum of the approximation (see (3.5))

$$\Phi^{(1)}(\tau, 1+\rho e^{i\beta\pi}) \sim \frac{\pi}{\sqrt{\tau}} \rho^{-\frac12}\Big[1 + \sqrt{2} C_\beta \Big(\frac{\tau}{\rho}\Big)^{\frac14}\Big]^{-2n} =: g^{(1)}(\tau,\rho), \tag{4.8}$$

where we have also used $|(1+\rho e^{i\beta\pi})^{-\frac12}| \sim \rho^{-\frac12}$. By solving

$$\frac{d}{d\rho}\rho^{-\frac12}\Big[1 + \sqrt{2} C_\beta \Big(\frac{\tau}{\rho}\Big)^{\frac14}\Big]^{-2n} = 0,$$

after some computations, we obtain

$$\hat\rho = 4C_\beta^4 \tau (n-1)^4 \cong \rho^\star. \tag{4.9}$$

4.1.2 Approximation of $\bar{\rho}$

For $\beta > \beta^\star$, the maximum of $\Phi^{(2)}(\tau, 1+\rho e^{-i\beta\pi})$ can be approximated by considering the ellipse $\Psi_{s_T} \in \mathcal{E}$ tangent to the curve $\chi^{(2)}(\Gamma_\beta^-)$. This is because the corresponding s_T represents the slowest convergence rate by (3.5). Let ρ_T be such that $\chi^{(2)}(1+\rho_T e^{-i\beta\pi})$ is the tangent point (Figure 3). Since the computation of ρ_T involves the solution of a fourth-degree equation, we consider an approximation arising from geometrical evidences. We first look for the ellipse $\Psi_{s_0} \in \mathcal{E}$ ($s_0 > s_T$) passing through the point $\chi^{(2)}(1) = \frac{2}{\sqrt{\tau}} i - 1$. Hence, we need to solve, with respect to s and φ,

$$\frac{1}{2}\left(s e^{i\varphi} + \frac{1}{s e^{i\varphi}}\right) = \frac{2}{\sqrt{\tau}} i - 1,$$

or equivalently,

$$\begin{cases} \frac{1}{2}\left(s\cos\varphi + \frac{1}{s}\cos\varphi\right) = -1, \\ \frac{1}{2}\left(s\sin\varphi - \frac{1}{s}\sin\varphi\right) = \frac{2}{\sqrt{\tau}}. \end{cases}$$

After some computations, we find that the solution (s_0, φ_0) is such that

$$\begin{aligned} \sin\varphi_0 &= \sqrt{\frac{2}{\tau}}\sqrt{\sqrt{1+\tau}-1}, \\ \cos\varphi_0 &= -\sqrt{1 - \frac{2}{\tau}(\sqrt{1+\tau}-1)}, \end{aligned} \tag{4.10}$$

with $\frac{\pi}{2} < \varphi_0 < \pi$, and

$$s_0 = \frac{\sqrt{2}}{\sqrt{\sqrt{1+\tau}-1}} + \frac{1}{\sqrt{1 - \frac{2}{\tau}(\sqrt{1+\tau}-1)}}. \tag{4.11}$$

The idea is to approximate $\bar{\rho}$ by looking for the other intersection between Ψ_{s_0} and $\chi^{(2)}(\Gamma_\beta^-)$ (see again Figure 3). In particular, we need to solve, with respect to ρ, the equation

$$\frac{1}{2}\left(s_0 e^{i\varphi} + \frac{1}{s_0 e^{i\varphi}}\right) = 2\left(\frac{1+\rho e^{-i\beta\pi}}{\tau}\right)^{\frac{1}{2}} i - 1.$$

Setting for simplicity $a = s_0 + \frac{1}{s_0}$ and $b = s_0 - \frac{1}{s_0}$, the above equation leads to the system

$$\begin{cases} \left(1 + \frac{1}{2}a\cos\varphi\right)^2 - \frac{1}{4}b^2\sin^2\varphi = -\frac{4}{\tau}\left(1 + \rho\cos(\beta\pi)\right), \\ b\sin\varphi\left(1 + \frac{1}{2}a\cos\varphi\right) = \frac{4}{\tau}\rho\sin(\beta\pi). \end{cases}$$

Substituting

$$\left(1 + \frac{1}{2}a\cos\varphi\right)^2 = \frac{16}{\tau^2}\frac{\rho^2\sin^2(\beta\pi)}{b^2\sin^2\varphi}$$

in the first equation, we obtain

$$\frac{16}{\tau^2}\frac{\rho^2\sin^2(\beta\pi)}{b^2\sin^2\varphi} - \frac{1}{4}b^2\sin^2\varphi + \frac{4}{\tau}(1+\rho\cos(\beta\pi)) = 0,$$

from which, after some computations, we find the solution

$$\tilde{\rho} = \frac{\tau b^2\sin^2\varphi}{8\sin^2(\beta\pi)}\left[-\cos(\beta\pi) \pm \sqrt{1 - \frac{16\sin^2(\beta\pi)}{\tau b^2\sin^2\varphi}}\right].$$

By using (4.10) and (4.11), we have that

$$b^2 = \left(s_0 - \frac{1}{s_0}\right)^2 = \frac{8}{\sqrt{1+\tau}-1} \sim \frac{8}{\sqrt{\tau}}, \quad \tau \to \infty.$$

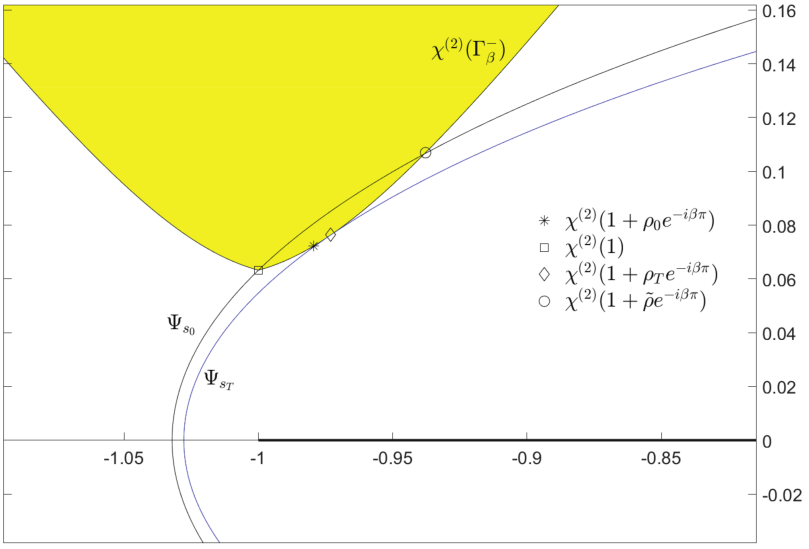

Figure 3: Geometrical interpretation of the local maximum of $\Phi^{(2)}(\tau, 1+\rho e^{-i\beta\pi})$ and its approximation by ρ_0 for $\beta = \frac{5}{12}$ and $\tau = 1000$.

Using again (4.7), we find

$$\tilde{\rho} \sim \frac{\tau^{\frac{1}{2}} \sin^2 \varphi}{\sin^2(\beta\pi)} \Big[-\cos(\beta\pi) \pm \Big(1 - \frac{\sin^2(\beta\pi)}{\tau^{\frac{1}{2}} \sin^2 \varphi} \Big) \Big].$$

Since the angle φ is still unknown and its computation requires again the solution of a fourth-degree equation, by taking the positive solution, we assume that (cf. (4.10))

$$\sin^2 \varphi \cong \sin^2 \varphi_0 \sim \frac{2}{\tau^{\frac{1}{2}}}$$

to finally obtain the rough approximation

$$\tilde{\rho} \cong \rho_0 := \frac{1 - \cos(\beta\pi)}{1 + \cos(\beta\pi)} = \tan^2\Big(\frac{\beta\pi}{2}\Big). \tag{4.12}$$

Experimentally, we observe that $0 \le \rho_0 < \tilde{\rho}$ and that $\chi^{(2)}(1+\rho_0 e^{-i\beta\pi})$ for large τ is close to the tangent point independently of β. Therefore, we use ρ_0 as an approximation of $\bar{\rho}$. As for the term $S^{(2)}$ in (3.6), we have the following result.

Proposition 2. *For $\rho = \rho_0$ and $\beta > \beta^\star$, we have*

$$S^{(2)} \sim 1 + \sqrt{2} G_\beta \tau^{-\frac{1}{4}}, \quad \tau \to \infty,$$

where

$$G_\beta = \sqrt{D_\beta - \sqrt{A_\beta^-}}, \tag{4.13}$$

with

$$D_\beta = (1 + 2\rho_0 \cos(\beta\pi) + \rho_0^2)^{\frac{1}{4}}, \tag{4.14}$$

$$A_\beta^- = \frac{-1 - \rho_0 \cos(\beta\pi) + \sqrt{1 + 2\rho_0 \cos(\beta\pi) + \rho_0^2}}{2}. \tag{4.15}$$

Proof. First of all, for $\rho = \rho_0$, we have that

$$t_0^{(2)} = \frac{2}{\sqrt{\tau}} (1 + \rho_0 \cos(\beta\pi) - i\rho_0 \sin(\beta\pi))^{\frac{1}{2}} - 1 = \frac{2}{\sqrt{\tau}} (i\sqrt{A_\beta^+} + \sqrt{A_\beta^-}) - 1,$$

where

$$A_\beta^+ = \frac{1 + \rho_0 \cos(\beta\pi) + \sqrt{1 + 2\rho_0 \cos(\beta\pi) + \rho_0^2}}{2}$$

and A_β^- as in (4.15). Now, defining $D_\beta := (1 + 2\rho_0 \cos(\beta\pi) + \rho_0^2)^{\frac{1}{4}}$ and using (4.7) for large τ,

$$\begin{aligned}
S^{(2)} &= \left| \frac{2}{\sqrt{\tau}}(i\sqrt{A_\beta^+} + \sqrt{A_\beta^-}) - 1 + \sqrt{\frac{4}{\tau}(A_\beta^- - A_\beta^+ + 2i\sqrt{A_\beta^- A_\beta^+}) - \frac{4}{\sqrt{\tau}}(\sqrt{A_\beta^-} + i\sqrt{A_\beta^+})} \right| \\
&\sim \left| -1 + \sqrt{-\frac{4}{\sqrt{\tau}}(\sqrt{A_\beta^-} + i\sqrt{A_\beta^+})} \right| \\
&= \left| -1 - \frac{\sqrt{2}}{\tau^{\frac{1}{4}}}\sqrt{-\sqrt{A_\beta^-} + D_\beta} + i\frac{\sqrt{2}}{\tau^{\frac{1}{4}}}\sqrt{\sqrt{A_\beta^-} + D_\beta} \right| \\
&= \sqrt{1 + \frac{2}{\tau^{\frac{1}{2}}}(-\sqrt{A_\beta^-} + D_\beta) + \frac{2\sqrt{2}}{\tau^{\frac{1}{4}}}\sqrt{-\sqrt{A_\beta^-} + D_\beta} + \frac{2}{\tau^{\frac{1}{2}}}(\sqrt{A_\beta^-} + D_\beta)} \\
&\sim \sqrt{1 + \frac{2\sqrt{2}}{\tau^{\frac{1}{4}}}\sqrt{D_\beta - \sqrt{A_\beta^-}}}.
\end{aligned}$$

Finally, using again (4.7) and defining

$$G_\beta := \sqrt{D_\beta - \sqrt{A_\beta^-}},$$

we obtain the result. □

We notice that $A_\beta^- = 0$, $D_\beta = 1$ and therefore $G_\beta = 1$ for $\beta = 0$. By using the above proposition and

$$|1 + \rho_0 e^{-i\beta\pi}| = \sqrt{1 + 2\rho_0 \cos(\beta\pi) + \rho_0^2} = D_\beta^2$$

(cf. (4.14)), we obtain (see (3.5))

$$\Phi^{(2)}(\tau, 1 + \rho_0 e^{-i\beta\pi}) \sim \frac{\pi}{\sqrt{\tau} D_\beta}(1 + \sqrt{2} G_\beta \tau^{-\frac{1}{4}})^{-2n} =: g^{(2)}(\tau, \rho_0). \tag{4.16}$$

4.2 The Optimal Value for τ

Working with approximations (4.8)–(4.9) and (4.16), in order to find a nearly optimal value for τ, we impose the condition

$$g^{(1)}(\tau, \hat{\rho}) = g^{(2)}(\tau, \rho_0). \tag{4.17}$$

Remark 2. For $0 \le \beta \le \beta^\star$, we should replace $g^{(2)}(\tau, \rho_0)$ by $g^{(2)}(\tau, 0)$ in (4.17). Anyway, since $\rho_0 \lesssim 0.17$ for $0 \le \beta \le \beta^\star \cong \frac{1}{4}$, we still work with (4.17) because this choice does not influence the results. Observe moreover that (4.12) luckily gives the correct value $\rho_0 = 0$ for $\beta = 0$.

Equation (4.17) leads to

$$\frac{\tau^{-\frac{1}{2}}}{2C_\beta^2 (n-1)^2}\left(1 + \frac{1}{n-1}\right)^{-2n} = \frac{1}{D_\beta}(1 + \sqrt{2} G_\beta \tau^{-\frac{1}{4}})^{-2n}.$$

Since

$$\left(1 + \frac{1}{n-1}\right)^{-2n} \sim e^{-2}, \tag{4.18}$$

after some computation, we rewrite the above equation as

$$\ln(\tau^{-\frac{1}{4}}) + \ln\left(\frac{D_\beta^{\frac{1}{2}}}{\sqrt{2} C_\beta e (n-1)}\right) = -n \ln(1 + \sqrt{2} G_\beta \tau^{-\frac{1}{4}}).$$

Let us denote by $W(x)$ the Lambert-W function, for which $\frac{x}{W(x)} = e^{W(x)}$ holds. By using

$$\ln(1 + \sqrt{2} G_\beta \tau^{-\frac{1}{4}}) \sim \sqrt{2} G_\beta \tau^{-\frac{1}{4}}, \quad \tau \gg 1,$$

we have that an approximated optimal value for τ is given by

$$\bar{\tau} = \frac{D_\beta^2}{4C_\beta^4 e^4 (n-1)^4} \exp(4W(H_\beta n(n-1))), \tag{4.19}$$

where

$$H_\beta := \frac{2eC_\beta G_\beta}{D_\beta^{\frac{1}{2}}}.$$

Note that, by (4.6), (4.13), (4.14), (4.15), we have $H_\beta = 2e$ for $\beta = 0$.

In Algorithm 1, we summarize the steps necessary to implement the method.

Algorithm 1. For any given n,
(1) set $\tau = \bar{\tau}$ by using (4.19) if $\alpha = \frac{1}{2}$ and empirically for $\alpha \neq \frac{1}{2}$;
(2) compute nodes and weights of the Gauss–Legendre rule;
(3) calculate the approximated value of $\mathcal{L}^{-\alpha}$ by using (2.4) and (2.5).

4.3 Asymptotic Expression of the Global Error

By substituting expression (4.19) in (4.9), we obtain

$$\hat{\rho} = \frac{D_\beta^2}{e^4} \exp(4W(H_\beta n(n-1))). \tag{4.20}$$

Since, for large x, $W(x) \sim \ln x - \ln(\ln x)$ (see [16]), we have that

$$\exp(4W(H_\beta n(n-1))) \sim \left[\frac{H_\beta n^2}{\ln(H_\beta n^2)}\right]^4,$$

and thus

$$\hat{\rho} \sim \frac{D_\beta^2}{e^4}\left[\frac{H_\beta n^2}{\ln(H_\beta n^2)}\right]^4. \tag{4.21}$$

Now, since $\hat{\rho}$ rapidly grows with n, as pointed out in Remark 1, we can estimate the global error as

$$\|E_n(\mathcal{L})\| \cong 4K\hat{\rho}^{-\frac{1}{2}} S^{(1)^{-2n}}, \tag{4.22}$$

where K is the absolute constant introduced in (4.1). Finally, substituting (4.21) in the above expression and using again approximation (4.18), after some computation, we find

$$\|E_n(\mathcal{L})\| \cong 4K\left[\frac{\ln(H_\beta n^2)}{2eC_\beta G_\beta}\right]^2 n^{-4}. \tag{4.23}$$

We remark that, in the above formula, the angle of the sector $\beta\pi$ asymptotically only affects the error constant so that the rate of convergence is independent of β.

Example 1. In order to test the behavior of the method, we consider a diagonal test matrix with a very large spectrum. In particular, we define

$$\mathcal{L} = \mathrm{diag}(1, 1+\rho_1 e^{i\beta\pi}, 1+\rho_1 e^{-i\beta\pi}, \ldots, 1+\rho_N e^{i\beta\pi}, 1+\rho_N e^{-i\beta\pi}),$$

where

$$\rho_i := 10^{x_i}, \quad x = (0, 0.1, 0.2, \ldots, 16)^T.$$

The square matrix $\mathcal{L}$ is of dimension $(2N+1)\times(2N+1)$, with $N = 161$, and it is clearly normal so that $F(\mathcal{L})$ is the convex hull of the spectrum, that is, the triangle with vertex at 1, $1 + 10^{16}e^{i\beta\pi}$, $1 + 10^{16}e^{-i\beta\pi}$. In Figure 4, we plot the error together with estimate (4.23) and the error of the sinc quadrature approximation implemented as in [7, formula (2) and Remark 3.1, with $s^+ = 0$] for some values of β. Here and below, we always consider the spectral norm of the error. In Table 1, we also show the corresponding values assumed by τ and $\hat{\rho}$ relative to the rightmost plot of Figure 4 ($\beta = \frac{5}{12}$).

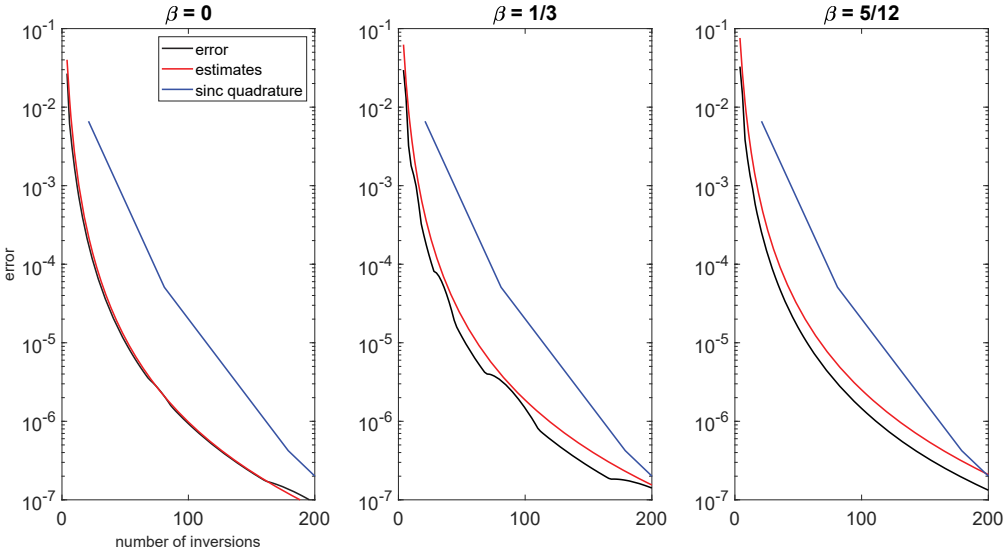

Figure 4: Example 1 – error, error estimate (4.23) and error of the sinc quadrature for the computation of $\mathcal{L}^{-\frac{1}{2}}$. Different values of β are considered.

n	10	25	40	55	70	85	100
τ	1.0E02	1.1E03	4.2E03	1.1E04	2.3E04	4.1E04	6.8E04
$\hat{\rho}$	4.1E05	2.2E08	6.0E09	5.7E10	3.1E11	1.3E12	4.0E12

Table 1: Example 1 – values of τ and $\hat{\rho}$ for $\beta = \frac{5}{12}$.

5 The Case of Bounded Operators

Consider the case of a bounded sectorial operator $\mathcal{L}_N$ with numerical range contained in

$$\Sigma_{\beta,1,\rho_N} = \{z \in \mathbb{C} : z = 1 + \rho e^{i\theta\pi},\ |\theta| \le \beta,\ 0 \le \rho \le \rho_N\}, \quad \beta < \frac{1}{2}.$$

Before starting, we need to remember that the local maximum $\rho^\star$ of $\Phi^{(1)}(\tau, 1+\rho e^{i\beta\pi})$ grows approximatively like $\frac{n^8}{(\ln n)^4}$ (see (4.20)). Therefore, there exists $\bar{n}$ such that $\rho^\star > \rho_N$ for $n > \bar{n}$. As a consequence, for $\rho^\star \le \rho_N$, estimate (4.23) is still valid because $\Phi^{(1)}(\tau, 1+\rho^\star e^{i\beta\pi}) \ge \Phi^{(1)}(\tau, 1+\rho_N e^{i\beta\pi})$, and hence we have to solve (4.17) to approximate the solution of (4.3). On the other side, for $\rho^\star > \rho_N$, the bound can be improved because $\rho^\star$ falls outside $[0, \rho_N]$. Similar to the analysis given in [2], for $\rho^\star > \rho_N$, the optimal value for τ can be approximated by solving

$$g^{(1)}(\tau, \rho_N) = g^{(2)}(\tau, \rho_0). \tag{5.1}$$

Proposition 3. *For $1 \ll \tau \ll \rho_N$, the solution of (5.1) is approximated by*

$$\hat{\tau} = \left(-\frac{\rho_N^{\frac{1}{4}}}{8\sqrt{2}C_\beta n} \ln\left(\frac{\sqrt{\rho_N}}{D_\beta}\right) + \sqrt{\left(\frac{\rho_N^{\frac{1}{4}}}{8\sqrt{2}C_\beta n} \ln\left(\frac{\sqrt{\rho_N}}{D_\beta}\right)\right)^2 + \frac{G_\beta}{C_\beta}\rho_N^{\frac{1}{4}}} \right)^4. \tag{5.2}$$

Proof. Using relations (4.8) and (4.16), equation (5.1) becomes

$$\rho_N^{-\frac{1}{2}} \left[1 + C_\beta\sqrt{2}\left(\frac{\tau}{\rho_N}\right)^{\frac{1}{4}}\right]^{-2n} = \frac{1}{D_\beta}(1 + \sqrt{2}G_\beta\tau^{-\frac{1}{4}})^{-2n}. \tag{5.3}$$

Using the approximations

$$1 + C_\beta\sqrt{2}\left(\frac{\tau}{\rho_N}\right)^{\frac{1}{4}} \cong \exp\left(C_\beta\sqrt{2}\left(\frac{\tau}{\rho_N}\right)^{\frac{1}{4}}\right) \tag{5.4}$$

and

$$1 + \sqrt{2}G_\beta\tau^{-\frac{1}{4}} \cong \exp(\sqrt{2}G_\beta\tau^{-\frac{1}{4}}),$$

we rewrite equation (5.3) as

$$\rho_N^{\frac{1}{4n}} \exp\Big(C_\beta\sqrt{2}\Big(\frac{\tau}{\rho_N}\Big)^{\frac{1}{4}}\Big) = D_\beta^{\frac{1}{2n}} \exp(G_\beta\sqrt{2}\tau^{-\frac{1}{4}}).$$

Therefore,

$$\tau^{\frac{1}{2}} + \frac{\rho_N^{\frac{1}{4}}}{4C_\beta\sqrt{2}n}\ln\Big(\frac{\sqrt{\rho_N}}{D_\beta}\Big)\tau^{\frac{1}{4}} - \frac{G_\beta}{C_\beta}\rho_N^{\frac{1}{4}} = 0.$$

By solving this equation and taking the positive solution, we obtain the result. □

We observe that, by (5.2), for $n \to +\infty$, we have

$$\begin{aligned}\Big(\frac{\hat\tau}{\rho_N}\Big)^{\frac{1}{4}} &= -\frac{1}{4\sqrt{2}C_\beta n}\ln\Big(\frac{\sqrt{\rho_N}}{D_\beta}\Big) + \sqrt{\Big(\frac{1}{4\sqrt{2}C_\beta n}\ln\Big(\frac{\sqrt{\rho_N}}{D_\beta}\Big)\Big)^2 + \frac{G_\beta}{C_\beta}\rho_N^{-\frac{1}{4}}}\\ &= -\frac{1}{4\sqrt{2}C_\beta n}\ln\Big(\frac{\sqrt{\rho_N}}{D_\beta}\Big) + \sqrt{\frac{G_\beta}{C_\beta}}\rho_N^{-\frac{1}{8}} + \mathcal{O}\Big(\frac{1}{n^2}\Big).\end{aligned}$$

Using (4.22), (4.5), (5.4) in order and the above result, for $n > \bar n$, we obtain

$$\begin{aligned}\|E_n(\mathcal{L}_N)\| &\cong 4\rho_N^{-\frac{1}{2}}\Big(1 + \sqrt{2}C_\beta\Big(\frac{\hat\tau}{\rho_N}\Big)^{\frac{1}{4}}\Big)^{-2n}\\ &\cong 4\rho_N^{-\frac{1}{2}}\exp\Big(-2\sqrt{2}C_\beta n\Big(\frac{\hat\tau}{\rho_N}\Big)^{\frac{1}{4}}\Big)\\ &\cong 4\rho_N^{-\frac{1}{4}}D_\beta^{-\frac{1}{2}}\exp(-2\sqrt{2}\sqrt{G_\beta C_\beta}n\rho_N^{-\frac{1}{8}}).\end{aligned} \tag{5.5}$$

Note that $\sqrt{G_\beta C_\beta} = 1$ for $\beta = 0$.

In order to derive an estimate of $\bar n$, we impose $\hat\rho = \rho_N$, where $\hat\rho$ is as in (4.20). We obtain the equation

$$\exp\big(W(2eC_\beta(n-1)n)\big) = e\rho_N^{\frac{1}{4}},$$

and therefore

$$W\big(2eC_\beta(n-1)n\big) = \ln(e\rho_N^{\frac{1}{4}}).$$

Since $W(z_1) = z_2$ if and only if $z_1 = z_2e^{z_2}$, it follows that

$$\bar n \cong \frac{\rho_N^{\frac{1}{8}}}{\sqrt{2C_\beta}}(\ln(e\rho_N^{\frac{1}{4}}))^{\frac{1}{2}} \cong \frac{\rho_N^{\frac{1}{8}}(\ln\rho_N)^{\frac{1}{2}}}{2\sqrt{2C_\beta}}. \tag{5.6}$$

In Algorithm 2, we summarize the steps necessary to implement the method in the case of a bounded operator.

Algorithm 2. For any given n,
(1) compute $\bar n$ as in (5.6);
(2) • $\alpha = \frac{1}{2}$: if $n < \bar n$, set $\tau = \bar\tau$ (see (4.19)); otherwise, set $\tau = \hat\tau$ (see (5.2));
  • $\alpha \neq \frac{1}{2}$: set τ empirically;
(3) compute nodes and weights of the Gauss–Legendre rule;
(4) calculate the approximated value of $\mathcal{L}_N^{-\alpha}$ by using (2.4) and (2.5).

Example 2. In order to test the method for sectorial bounded operators, we first consider the same operator of Example 1 but with

$$x = (0, 0.1, 0.2, \dots, 4)^T$$

so that $\rho_N = 10^4$ and $N = 41$. In Figure 5, the error and estimate (5.5) are plotted for different values of β.

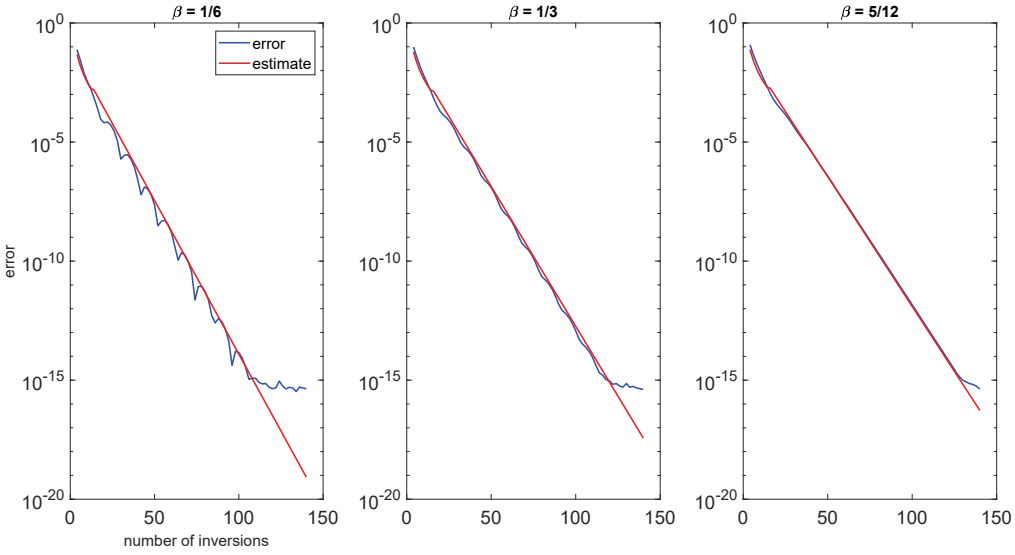

Figure 5: Example 2 – error and error estimate (5.5) for the computation of $\mathcal{L}^{-\frac{1}{2}}$. Different values of β are considered.

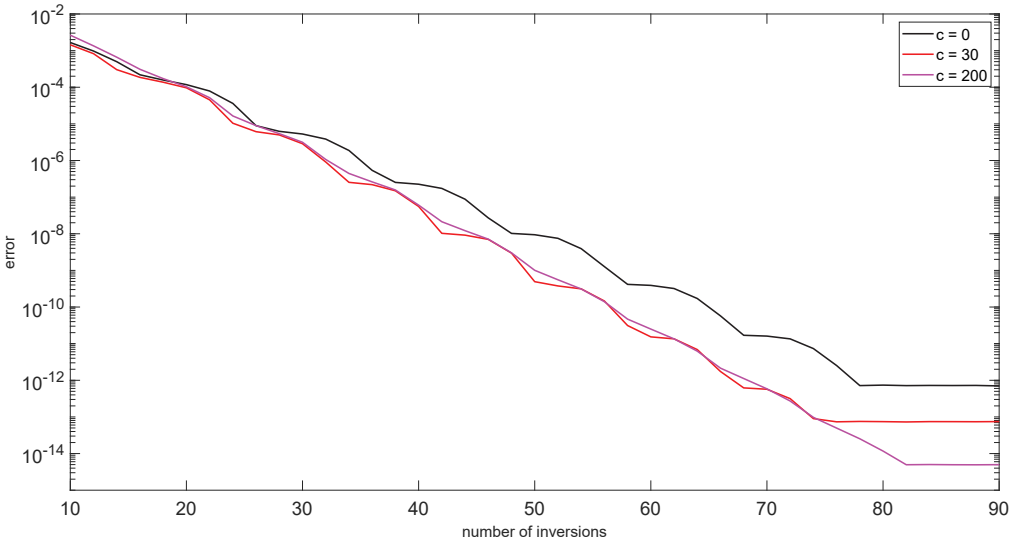

Figure 6: Example 3 – error for the computation of $\mathcal{L}_N^{-\frac{1}{2}}$.

Example 3. As a more realistic example, we also consider the discretization using central differences of the operator

$$\mathcal{L}u = -u'' + cu', \quad c \geq 0,$$

on $[0, 1]$ with Dirichlet boundary conditions. We have taken $N = 200$ equally spaced interior points. By moving the constant c, we change the angle $\beta\pi$ of the sector containing $F(\mathcal{L}_N)$, where $\mathcal{L}_N$ is the discretization matrix. In Figure 6, we plot the errors for $c = 0$ ($\beta = 0$), $c = 30$ ($\beta = 0.44$) and $c = 200$ ($\beta = 0.49$). It is interesting to observe that the method is a bit faster for $c \gg 0$. This is due to the position of the eigenvalues of smallest modulus that move away from 0 for growing c. With large c, we also notice an improvement of the attainable accuracy, and the reason lies in the conditioning of $\mathcal{L}_N$ that reduces by increasing c.

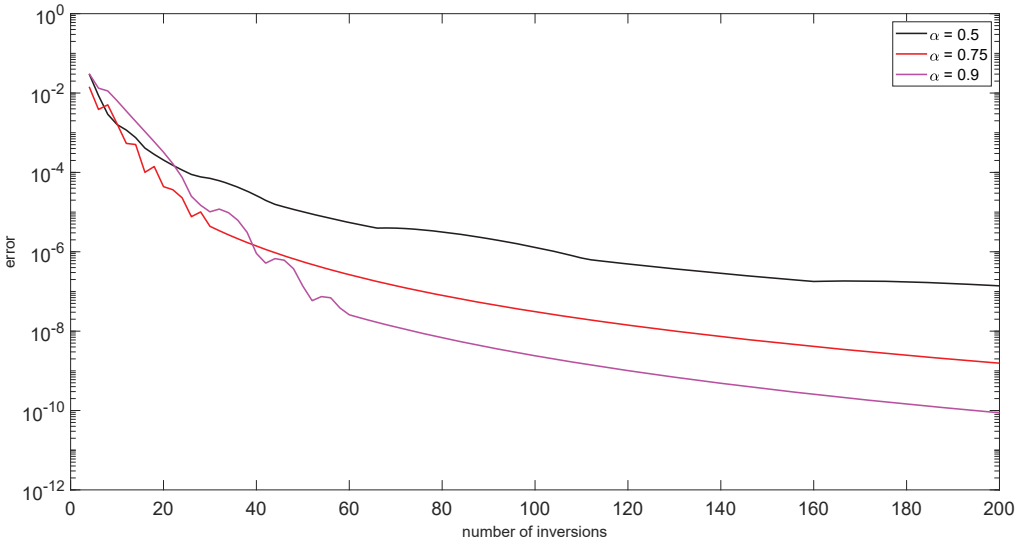

Figure 7: Error for the computation of $\mathcal{L}^{-\alpha}$, $\mathcal{L}$ as in Example 1 with $\beta = \frac{1}{6}$, for different values of α.

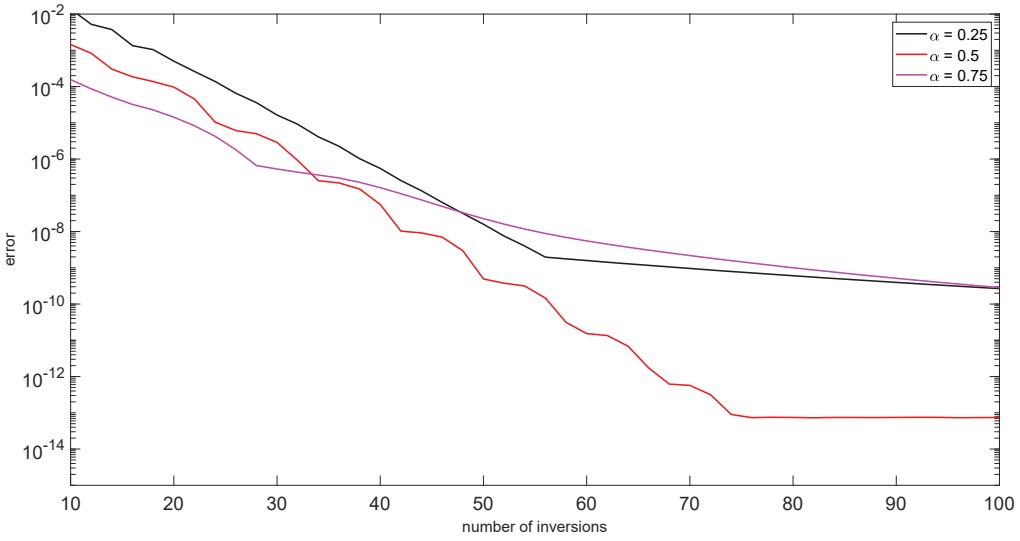

Figure 8: Error for the computation of $\mathcal{L}_N^{-\alpha}$, $\mathcal{L}_N$ as in Example 3 with $c = 30$, for different values of α.

6 Experiments for $\alpha \neq \frac{1}{2}$

While all the analysis is restricted to the case of $\alpha = \frac{1}{2}$, we remark that the method works rather fine also more generally for $\mathcal{L}^{-\alpha}$, $0 < \alpha < 1$. Working with the operator defined in Example 1, in Figure 7, we compare the behavior of the method for $\alpha = 0.75$ and $\alpha = 0.9$ with respect to the case $\alpha = 0.5$, using the value of τ given by expression (4.19). Unfortunately, this choice does not work well for $\alpha < 0.5$.

Moreover, in Figure 8, we consider the operator of Example 3, with $c = 30$, and apply the method also for $\alpha = 0.25$ and $\alpha = 0.75$. By Algorithm 2, we have empirically set $\tau = \frac{\bar{\tau}}{5}$ ($\tau = \frac{\hat{\tau}}{5}$) for $\alpha = 0.25$ and $\tau = \frac{3\bar{\tau}}{2}$ ($\tau = \frac{3\hat{\tau}}{2}$) for $\alpha = 0.75$, where $\bar{\tau}$ is defined in (4.19) and $\hat{\tau}$ in (5.2). We observe that, for $\alpha \neq 0.5$, the error initially decreases with a rate of convergence similar to the case $\alpha = 0.5$ and then shows a progressively slow down. This is due to the fact that, since for $\alpha \neq 0.5$, we do not have accurate error estimates, we are not able to suitable define τ. Anyway, as a general indication, one should take $\tau = c\bar{\tau}$ ($\tau = c\hat{\tau}$), with $c < 1$ for $\alpha < 0.5$ and $c > 1$ for $\alpha > 0.5$. We point out, however, that, even without an optimal choice, the method appears to be initially quite fast (Figure 8) so that it can be fruitfully used in applications requiring moderate accuracy.

7 Conclusions

We have studied an unexplored Gaussian approach for the computation of the inverse square root of regularly accretive operators. The method exhibits a very fast initial convergence, and its rate is independent of the angle of the sector containing the numerical range. We have derived sharp error estimates that can be used for an a priori selection of the number of quadrature points necessary to achieve a prescribed accuracy. We have shown that the method can also be used for a generic fractional power, but in this general situation, the parameter τ needs to be defined empirically.

A Approximation of $\beta^\star$

Let $\Psi_{s_0} \in \mathcal{E}$ be the ellipse passing through the point $\frac{2}{\sqrt{\tau}}i - 1$ (cf. (4.11)). The value $\beta^\star$, such that the function $\Phi^{(2)}(1 + \rho e^{-i\beta\pi})$ possesses a local maximum for $\beta > \beta^\star$, is the one for which Ψ_{s_0} is also tangent to the curve $\chi^{(2)}(\Gamma^-_{\beta^\star})$ at $\frac{2}{\sqrt{\tau}}i - 1$ (Figure 3). In order to compute $\beta^\star$, we consider the tangents at $\frac{2}{\sqrt{\tau}}i - 1$ to the ellipse and to the curve, and impose them to have the same slope. Before starting, we need to derive the semi-width γ and the semi-height δ of the ellipse. By geometrical evidence, we have that

$$\delta = \mathbb{I}\left\{\frac{1}{2}\left(s_0 e^{i\frac{\pi}{2}} + \frac{1}{s_0}e^{-i\frac{\pi}{2}}\right)\right\} = \frac{1}{2}\left(s_0 - \frac{1}{s_0}\right) \tag{A.1}$$

and $\gamma^2 = \delta^2 + 1$. At this point, we remind that the slope of the tangent at $\frac{2}{\sqrt{\tau}}i - 1$ to the ellipse is

$$m = -\frac{\delta^2}{\gamma^2}\frac{\mathbb{R}(\frac{2}{\sqrt{\tau}}i - 1)}{\mathbb{I}(\frac{2}{\sqrt{\tau}}i - 1)} = \frac{\delta^2}{\delta^2 + 1}\left(\frac{\sqrt{\tau}}{2}\right) \sim \frac{\sqrt{\tau}}{2 + \sqrt{\tau}} \sim 1, \quad \tau \to +\infty, \tag{A.2}$$

where we have used (A.1) and

$$s_0 - \frac{1}{s_0} \sim \frac{4}{\sqrt{2}}\tau^{-\frac{1}{4}},$$

which comes from (4.11). Now, it is not difficult to show that the angle between the tangent to the curve $\chi^{(2)}(\Gamma^-_\beta)$ at $\frac{2}{\sqrt{\tau}}i - 1$ and the line $\mathbb{I}(z) = \frac{2}{\sqrt{\tau}}$ is given by $\frac{1}{2}(\pi - 2\beta\pi)$. Hence, in order to find an approximation of $\beta^\star$, by (A.2), we impose the condition

$$\tan\left[\frac{\pi}{2}(1 - 2\beta)\right] = 1$$

that leads to $\beta^\star \sim \frac{1}{4}$ as $\tau \to +\infty$.

Funding: This work was partially supported by GNCS-INdAM, FRA-University of Trieste and CINECA under HPC-TRES program award number 2019-04. The authors are members of the INdAM research group GNCS.

References

[1] L. Aceto, D. Bertaccini, F. Durastante and P. Novati, Rational Krylov methods for functions of matrices with applications to fractional partial differential equations, *J. Comput. Phys.* **396** (2019), 470–482.

[2] L. Aceto and P. Novati, Rational approximations to fractional powers of self-adjoint positive operators, *Numer. Math.* **143** (2019), no. 1, 1–16.

[3] L. Aceto and P. Novati, Padé-type approximations to the resolvent of fractional powers of operators, *J. Sci. Comput.* **83** (2020), no. 1, Paper No. 13.

[4] L. Aceto and P. Novati, Fast and accurate approximations to fractional powers of operators, *IMA J. Numer. Anal.* **42** (2022), no. 2, 1598–1622.

[5] A. V. Balakrishnan, Fractional powers of closed operators and the semigroups generated by them, *Pacific J. Math.* **10** (1960), 419–437.

[6] W. Barrett, Convergence properties of Gaussian quadrature formulae, *Comput. J.* **3** (1960/61), 272–277.
[7] A. Bonito, W. Lei and J. E. Pasciak, On sinc quadrature approximations of fractional powers of regularly accretive operators, *J. Numer. Math.* **27** (2019), no. 2, 57–68.
[8] A. Bonito and J. E. Pasciak, Numerical approximation of fractional powers of elliptic operators, *Math. Comp.* **84** (2015), no. 295, 2083–2110.
[9] M. Crouzeix and C. Palencia, The numerical range is a $(1 + \sqrt{2})$-spectral set, *SIAM J. Matrix Anal. Appl.* **38** (2017), no. 2, 649–655.
[10] S. Harizanov, R. Lazarov and S. Margenov, A survey on numerical methods for spectral space-fractional diffusion problems, *Fract. Calc. Appl. Anal.* **23** (2020), no. 6, 1605–1646.
[11] S. Harizanov, R. Lazarov, S. Margenov, P. Marinov and J. Pasciak, Comparison analysis of two numerical methods for fractional diffusion problems based on the best rational approximations of t^γ on $[0, 1]$, in: *Advanced Finite Element Methods with Applications*, Lect. Notes Comput. Sci. Eng. 128, Springer, Cham (2019), 165–185.
[12] S. Harizanov, R. Lazarov, S. Margenov, P. Marinov and J. Pasciak, Analysis of numerical methods for spectral fractional elliptic equations based on the best uniform rational approximation, *J. Comput. Phys.* **408** (2020), Article ID 109285.
[13] S. Harizanov, R. Lazarov, S. Margenov, P. Marinov and Y. Vutov, Optimal solvers for linear systems with fractional powers of sparse SPD matrices, *Numer. Linear Algebra Appl.* **25** (2018), no. 5, Article ID e2167.
[14] S. Harizanov and S. Margenov, Positive approximations of the inverse of fractional powers of SPD M-matrices, in: *Control Systems and Mathematical Methods in Economics*, Lecture Notes Econom. Math. Systems 687, Springer, Cham (2018), 147–163.
[15] C. Hofreither, A unified view of some numerical methods for fractional diffusion, *Comput. Math. Appl.* **80** (2020), no. 2, 332–350.
[16] A. Hoorfar and M. Hassani, Inequalities on the Lambert *W* function and hyperpower function, *JIPAM. J. Inequal. Pure Appl. Math.* **9** (2008), no. 2, Article ID 51.
[17] T. Kato, Fractional powers of dissipative operators, *J. Math. Soc. Japan* **13** (1961), 246–274.
[18] L. N. Trefethen, Is Gauss quadrature better than Clenshaw–Curtis?, *SIAM Rev.* **50** (2008), no. 1, 67–87.
[19] P. N. Vabishchevich, Numerically solving an equation for fractional powers of elliptic operators, *J. Comput. Phys.* **282** (2015), 289–302.
[20] P. N. Vabishchevich, Numerical solution of time-dependent problems with fractional power elliptic operator, *Comput. Methods Appl. Math.* **18** (2018), no. 1, 111–128.
[21] P. N. Vabishchevich, Approximation of a fractional power of an elliptic operator, *Numer. Linear Algebra Appl.* **27** (2020), no. 3, Article ID e2287.