
Computing the Inverse of a ϕ -Function by Rational Approximation

Paola Boito · Yuli Eidelman · Luca Gemignani

the date of receipt and acceptance should be inserted later

Abstract In this paper we introduce a family of rational approximations of the inverse of a ϕ function involved in the explicit solutions of certain linear differential equations as well as in integration schemes evolving on manifolds. For symmetric banded matrices these novel approximations provide a computable reconstruction of the associated matrix function which exhibits decaying properties comparable to the best existing theoretical bounds. Numerical examples show the benefits of the proposed rational approximations w.r.t. the classical Taylor polynomials.

Keywords rational approximation · ϕ -functions · band matrices · matrix functions

Mathematics Subject Classification (2000) 65F60

1 Introduction

Numerical methods for the computation of matrix functions involved in the solution of certain ordinary or partial differential equations have witnessed growing interest in recent years (see [11] and the references given therein). Several methods are concerned with the evaluation of the matrix ϕ -functions $\phi_k(A)$, $k \geq 0$, for a large and possibly sparse matrix A , where $\phi_k(z)$ are entire functions defined recursively by $\phi_{k+1}(z) = \frac{\phi_k(z) - (1/k!)}{z}$ with $\phi_0(z) = e^z$ [12]. Here we focus on the related issue of approximating the inverse ϕ_1 matrix function $\psi_1(A)$ where $\psi_1(z)$ is a meromorphic function defined by

$$\psi_1(z) = \phi_1(z)^{-1} = \frac{z}{e^z - 1}$$

P. Boito
Dipartimento di Matematica, Università di Pisa, Largo Bruno Pontecorvo, 5 - 56127 Pisa, Italy. E-mail: paola.boito@unipi.it

Y. Eidelman
School of Mathematical Sciences, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel-Aviv University, Ramat-Aviv, 69978, Israel. E-mail: eideyu@post.tau.ac.il

L. Gemignani
Dipartimento di Informatica, Università di Pisa, Largo Bruno Pontecorvo, 3 - 56127 Pisa, Italy. E-mail: luca.gemignani@unipi.it

and A is banded or more generally rank-structured (see [8] for a survey on such matrices). This problem also plays an important role in a number of applications.

Two-point inverse problems for first order differential equations are frequently encountered in mathematical physics (see Chapter 7 in [18]). Let us consider the differential problem

$$\frac{d\mathbf{u}}{dt} = A\mathbf{u}(t) + \mathbf{p}, \quad 0 \leq t \leq \tau, \quad (1.1)$$

where $A \in \mathbb{R}^{d \times d}$ is given, while $\mathbf{p} \in \mathbb{R}^d$ is unknown. In order to find the solution $\mathbf{u}: [0, \tau] \rightarrow \mathbb{R}^d$ of (1.1), and the vector \mathbf{p} simultaneously the overdetermined conditions

$$\mathbf{u}(0) = \mathbf{u}_0 = \mathbf{g}, \quad \mathbf{u}(\tau) = \mathbf{h}, \quad (1.2)$$

can be imposed. A more general formulation of the inverse problem (1.1), (1.2) in a Banach space with a closed linear operator A is treated in [22, 21]. A new formula for the solution of the problem (1.1), (1.2) using Bernoulli polynomials is given in [19].

Assume that the complex numbers

$$2\pi ik/\tau, \quad k = \pm 1, \pm 2, \dots \quad (1.3)$$

do not belong to the spectrum of A . Define the complex-valued functions

$$q_t(z) = \frac{z}{e^{\tau z} - 1} e^{zt}, \quad w_t(z) = \frac{e^{zt} - 1}{e^{\tau z} - 1}, \quad 0 \leq t \leq \tau, \quad z \in \mathbb{C} \quad (1.4)$$

with $q_t(0) = \frac{1}{\tau}$, $w_t(0) = \frac{t}{\tau}$. The complex functions $q_t(z), w_t(z)$ are meromorphic in z with the poles (1.3). One can check directly that the solution of the inverse problem (1.1), (1.2) is given by the formulas

$$\mathbf{p} = q_0(A)(\mathbf{h} - \mathbf{g}) - A\mathbf{g} \quad (1.5)$$

and

$$\mathbf{u}(t) = w_t(A)(\mathbf{h} - \mathbf{g}) + \mathbf{g}, \quad 0 \leq t \leq \tau. \quad (1.6)$$

Using (1.5) and the formula $q_0(z) = \psi_1(\tau z)/\tau$ we obtain the formula

$$\mathbf{p} = \frac{1}{\tau} \psi_1(\tau A)(\mathbf{h} - \mathbf{g}) - A\mathbf{g} \quad (1.7)$$

to compute the unknown vector \mathbf{p} via the function ψ_1 .

Notice also that the formula

$$\mathbf{v}(t) = q_t(A)\mathbf{v}_0, \quad 0 \leq t \leq \tau$$

yields the solution of the nonlocal problem

$$\frac{d\mathbf{v}}{dt} = A\mathbf{v}(t), \quad 0 \leq t \leq \tau, \quad \int_0^\tau \mathbf{v}(t) dt = \mathbf{v}_0$$

studied by the authors in [4].

Computing the inverse of $\phi_1(B)$, $B \in \mathbb{R}^{d \times d}$ also plays a fundamental role in the application of exponential integrators for the numerical solution of systems of differential equations. The reason is twofold. First, certain integration schemes called Runge-Kutta Munthe-Kaas methods [17, 16, 13] for computing numerical solutions

of differential equations that are guaranteed to evolve on a prescribed manifold require explicitly the approximation of the inverse of the linear map $g: \mathbb{R}^{d \times d} \rightarrow \mathbb{R}^{d \times d}$ given by $g(B) = \phi_1([A, B]) = (e^{[A, B]} - I_d)([A, B])^{-1}$ where $A \in \mathbb{R}^{d \times d}$ is fixed and $[A, B] = AB - BA$ is the matrix commutator. Including the evaluation of the map $\psi_1([A, B])$ in numerical algorithms seems to be awkward and several polynomial approximations of $\psi_1(z)$ have been presented in the related literature (compare with [5] and the references given therein). Secondly, the study of reliable procedures for the evaluation of $\psi_1(A)$ and $\psi_1(A)\mathbf{v}$ based on rational approximations of the meromorphic function $\psi_1(z)$ might be used to foster the development of rational Krylov methods for computing $\phi_1(A)$ and $\phi_1(A)\mathbf{v}$ that is the main computational bulk in exponential integrators for stiff systems of differential equations [9]. Indeed the properties of these methods depend heavily on the features of the underlying rational approximations for the selection of the poles and of the subspace of approximation.

Customary approximations of $\psi_1(B)$ derived from truncated Taylor expansions goes back to the work of Magnus [14]. These approximations are quite accurate if the norm of the matrix B is sufficiently small. On the other hand, rational functions may exhibit approximation properties and convergence domains superior to polynomials provided that the poles of the rational functions involved have been chosen in a suitable way. Moreover, if B is banded or even just rank-structured then the same property holds in a certain approximate sense for the matrix $\phi_1(B)$ and thus a fortiori for its inverse $\psi_1(B)$. Polynomial approximations for the function $\psi_1(z)$ often require a quite high degree of the approximating polynomial in order to achieve a reasonable quality of approximation of the numerical rank structure and the decaying properties of the matrix $\psi_1(B)$. Rational approximations would typically obtain the same quality with substantially fewer degrees of freedom.

In this paper we present new algorithms that efficiently approximate the functions of a matrix argument involved in the solution of (1.1),(1.2). In particular, we propose a novel family of mixed polynomial-rational approximations of $\psi_1(z)$ required for the computation of the vector \mathbf{p} according to (1.7). By combining Fourier analysis methods applied to the function $q_t(z)/z$ with classical tools for Fourier series acceleration [7] for any fixed $s > 1$ and $m \geq s$ we obtain approximations of $\psi_1(B) = (\phi_1(B))^{-1}$ of the form

$$\psi_1(B) \simeq p_s(B) + \sum_{k=1}^m \gamma_{k,s} B^{\tau_s} (B^2 + k^2 I_d)^{-1}, \quad (1.8)$$

where $p_s(z)$ is a polynomial of degree $\ell = \ell(s)$ and $\tau_s = \tau(s) \in \mathbb{N}$. These novel expansions have better theoretical and computational properties than polynomial approximations based on the Maclaurin series. In particular numerical evidence suggests that the formulas (1.8) are accurate on larger domains thus allowing for larger integration steps.

Two potential drawbacks against the use of rational formulas are the possible higher computational cost of linear system solving w.r.t. matrix-by-vector multiplication and the requirement of the full storage of the matrix for direct solvers. However, these issues do not apply for remarkable classes of matrices including band, rank structured and displacement structured matrices. It is shown that for a symmetric banded matrix A these novel approximations (1.8) yield a computable reconstruction of the associated matrix function $\psi_1(A)$ which exhibits

decaying properties comparable to the best existing theoretical bounds and significantly superior to the behavior of the corresponding polynomial approximations. The matrix function $\psi_1(A)$ can thus be manipulated efficiently using its resulting data-sparse format combined with the rank-structured matrix technology [8]. Furthermore, when A is rank-structured the action of the matrix $\psi_1(A)$ on a vector can be computed efficiently using the direct fast solver for shifted linear systems proposed in [4].

The paper is organized as follows. In section 2 we present a general scheme for the design of accurate rational approximations of the matrix functions involved in the solution of (1.1),(1.2). In Subsection 2.1 this scheme is specialized for the construction of mixed polynomial-rational approximations of the meromorphic function $\psi_1(z)$. In section 3 we investigate both theoretical and computational properties of the application of these formulas to computing $\psi_1(A)$ where A is a symmetric banded matrix. Finally, conclusions and future work are presented in section 4.

2 Rational Approximation of the Inverse Problem and the Inverse ϕ_1 -Function

The solvability of the inverse problem (1.1), (1.2) in an abstract Banach space is studied in [22,21,19]. Under the assumption that all the numbers $\mu_k = 2\pi ik/\tau$, $k = \pm 1, \pm 2, \pm 3, \dots$ are regular points of the linear operator A the inverse problem (1.1), (1.2) has a unique solution. Without loss of generality one can assume that $\tau = 2\pi$. As it was mentioned above in the matrix case this solution is given by the formulas (1.4), (1.5), (1.6).

We introduce the auxiliary function

$$r_t(z) = \frac{e^{zt}}{e^{z2\pi} - 1}, \quad 0 \leq t \leq 2\pi. \quad (2.9)$$

Using the formulas (1.4) we have

$$q_t(z) = zr_t(z), \quad w_t(z) = r_t(z) - r_0(z), \quad 0 \leq t \leq 2\pi. \quad (2.10)$$

Expanding the function $r_t(z)$ in the Fourier series of t we obtain

$$r_t(z) = \frac{1}{2\pi} \left(\frac{1}{z} + \sum_{k \in \mathbb{Z}/\{0\}} \frac{e^{ikt}}{z - ik} \right), \quad 0 < t < 2\pi.$$

We consider the equivalent representation with the real series given by

$$r_t(z) = \frac{1}{2\pi z} + \frac{1}{\pi} \sum_{k=1}^{\infty} (z \cos(kt) - k \sin(kt))(z^2 + k^2)^{-1}, \quad 0 < t < 2\pi. \quad (2.11)$$

It is well known that the convergence of this series must depend strongly on the smoothness of the periodic extension of $r_t(z)$. Acceleration techniques proposed in [7] make use of the Bernoulli polynomials for the approximate reconstruction of jumps.

Applying the formula

$$\frac{1}{z^2 + k^2} = \frac{1}{k^2} - \frac{z^2}{k^2(z^2 + k^2)} \quad (2.12)$$

to the last entry in (2.11) we obtain that for $0 < t < 2\pi$ it holds

$$r_t(z) = \frac{1}{2\pi z} + \frac{1}{\pi} \sum_{k=1}^{\infty} (z \cos(kt) + \frac{1}{k} z^2 \sin(kt))(z^2 + k^2)^{-1} - \frac{1}{\pi} \sum_{k=1}^{\infty} \frac{1}{k} \sin(kt).$$

Since

$$2 \sum_{k=1}^{\infty} \frac{1}{k} \sin kt = \pi - t, \quad 0 < t < 2\pi$$

we arrive at the following formula for $0 < t < 2\pi$,

$$r_t(z) = \frac{1}{2\pi z} + \frac{t - \pi}{2\pi} + y_t(z) \quad (2.13)$$

with

$$y_t(z) = \frac{1}{\pi} \sum_{k=1}^{\infty} (z \cos(kt) + \frac{1}{k} z^2 \sin(kt))(z^2 + k^2)^{-1}. \quad (2.14)$$

If $z \in K \subset \mathbb{C}$, K compact set, then definitively we have

$$|z - ik|^{-1} \leq \frac{C}{|k|}, \quad (2.15)$$

and, therefore, using the Weierstrass M-test one can easily check that the series in (2.13) converges uniformly in $(t, z) \in [0, 2\pi] \times K$. Hence, by continuity we may extend the formula (2.14) over the whole interval $[0, 2\pi]$.

Combining the formulas (1.4) and (2.13) we get

$$q_t(z) = \frac{1}{2\pi} + z \frac{t - \pi}{2\pi} + zy_t(z) \quad (2.16)$$

and

$$w_t(z) = \frac{t}{2\pi} + (y_t(z) - y_0(z)). \quad (2.17)$$

Here there are no singularities at $z = 0$. Inserting (2.16), (2.17) in (1.5), (1.6) we obtain

$$\mathbf{p} = \left(\frac{1}{2\pi} I - \frac{1}{2} A + Ay_0(A) \right) (\mathbf{h} - \mathbf{g}) - A\mathbf{g} \quad (2.18)$$

and

$$\mathbf{u}(t) = \left(\frac{t}{2\pi} I + (y_t(A) - y_0(A)) \right) (\mathbf{h} - \mathbf{g}) + \mathbf{g}, \quad 0 \leq t \leq 2\pi \quad (2.19)$$

with

$$y_t(A) = \frac{1}{\pi} \sum_{k=1}^{\infty} (A \cos(kt) + \frac{1}{k} A^2 \sin(kt))(A^2 + k^2 I)^{-1}. \quad (2.20)$$

The rate of convergence of the series in (2.20) is the same as for the series $\sum_{k=1}^{\infty} k^{-2}$. The last may be improved by using repeatedly the equality (2.12) as

above. For each integer $k \geq 0$ denote as $B_k(t)$ the Bernoulli polynomials (extended by periodicity onto the real line) defined by

$$\frac{ze^{zt}}{e^z - 1} = \sum_{k=0}^{+\infty} B_k(t) \frac{z^k}{k!}, \quad |t| < 2\pi, \quad (2.21)$$

$$B_k = B_k(0), \quad k \geq 0,$$

where B_k are the Bernoulli numbers. Then, using (2.12) we prove by induction the following formulas.

Lemma 1 *Let $t \in [0, 2\pi]$ and $y_t(z) : \mathbb{C} \rightarrow \mathbb{C}$ be defined as in (2.14). Then we have*

$$y_t(z) = p_{n,t}(z) + s_{n,t}(z), \quad n = 0, 1, 2, \dots \quad (2.22)$$

with

$$p_{n,t}(z) = \sum_{i=2}^{2n+1} \frac{(2\pi)^{i-1}}{i!} B_i\left(\frac{t}{2\pi}\right) z^{i-1} \quad (2.23)$$

and

$$s_{n,t}(z) = \frac{(-1)^n}{\pi} \sum_{k=1}^{\infty} \frac{z^{2n}(z \cos(kt) + \frac{1}{k} z^2 \sin(kt))}{k^{2n}(z^2 + k^2)}. \quad (2.24)$$

Proof For $n = 0$ the relation (2.22) follows directly from (2.14). Assume that for some $n \geq 0$ the relation (2.22) holds. Using (2.12) we have

$$\begin{aligned} s_{n,t}(z) &= \frac{(-1)^{n+1}}{\pi} \sum_{k=1}^{\infty} \frac{z^{2(n+1)}(z \cos(kt) + \frac{1}{k} z^2 \sin(kt))}{k^{2(n+1)}(z^2 + k^2)} + \\ &\quad \frac{(-1)^n}{\pi} \sum_{k=1}^{\infty} \frac{z^{2n}(z \cos(kt) + \frac{1}{k} z^2 \sin(kt))}{k^{2(n+1)}}, \end{aligned}$$

i.e.

$$s_{n,t}(z) = s_{n+1,t}(z) + b_{n,t}(z), \quad (2.25)$$

where

$$b_{n,t}(z) = \frac{(-1)^n}{\pi} z^{2n+1} \sum_{k=1}^{\infty} \frac{\cos(kt)}{k^{2(n+1)}} + \frac{(-1)^n}{\pi} z^{2n+2} \sum_{k=1}^{\infty} \frac{\sin(kt)}{k^{2n+3}}.$$

Using the formulas from [[1], formula 23.1.18] we know that

$$\sum_{k=1}^{\infty} \frac{\cos(kt)}{k^{2(n+1)}} = \frac{(2\pi)^{2(n+1)} B_{2(n+1)}(t/(2\pi))}{(-1)^n 2(2(n+1))!}$$

and

$$\sum_{k=1}^{\infty} \frac{\sin(kt)}{k^{2n+3}} = \frac{(2\pi)^{2n+3} B_{2(n+1)}(t/(2\pi))}{(-1)^n 2(2n+3)!}.$$

Hence it follows that

$$b_{n,t}(z) = \frac{z^{2n+1} (2\pi)^{2n+1} B_{2n+2}(t/(2\pi))}{(2n+2)!} + \frac{z^{2n+2} (2\pi)^{2n+2} B_{2n+3}(t/(2\pi))}{(2n+3)!}.$$

Inserting this in (2.25) and using (2.22) we complete the proof of the statement.

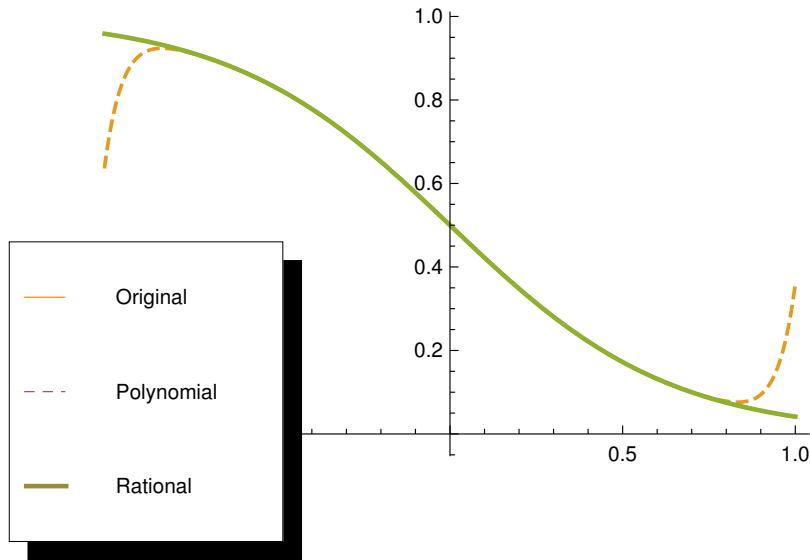


Fig. 1: Polynomial and rational approximations of $w_t(z)$. Note that the plots of $w_t(z)$ and of the rational approximation overlap.

Based on this result in the next section we derive accelerated series representation of the function $y_t(z)$ and a fortiori of $q_t(z)$ and $\psi_1(z)$. In Figure 1 we give an illustration of the quality of the rational approximation of the function $w_t(z)$ obtained from (2.17) combined with the decompositions introduced in Lemma 1. More precisely we consider the (original) function $w_\pi(z)$ and compare the accuracy of the (polynomial) approximation of degree 19 generated from (2.21) and the (rational) approximation derived from Lemma 1 with $n = 4$ and by truncating the series to the first 10 terms.

2.1 The application to the ψ_1 -function

It follows that $q_0(z) = \psi_1(2\pi z)/(2\pi)$ and $\psi_1(z)$ admits a Maclaurin series expansion which can virtually be used to evaluate $q_0(A)$. The following classical result provides the Maclaurin expansion of $\psi_1(z)$.

Theorem 1 ([1], formula 23.1.1) *It holds*

$$\psi_1(z) = \phi_1(z)^{-1} = \sum_{k=0}^{+\infty} \frac{B_k}{k!} z^k, \quad |z| < 2\pi,$$

where B_k denotes the k th Bernoulli number.

Different rational approximations of $\psi_1(z)$ can be derived from the Fourier series expansion of $r_t(z)$. It turns out that such series representation is also related with the Mittag-Leffler expansion of $q_0(z)$.

Specifying the formulas and representations obtained above to the function $\psi_1(A)$ we obtain the following. Using the formula (2.16) we have

$$q_t(A) = \frac{1}{2\pi}I + \frac{t-\pi}{2\pi}A + Ay_t(A) \quad (2.26)$$

and using (2.20) we find that

$$\begin{aligned} q_t(A) &= \frac{1}{2\pi}I_d + \frac{t-\pi}{2\pi}A + \frac{1}{\pi} \sum_{k=1}^{\infty} A^2 \cos(kt)(A^2 + k^2I_d)^{-1} + \\ &+ \frac{1}{\pi} \sum_{k=1}^{\infty} \frac{1}{k} A^3 \sin(kt)(A^2 + k^2I_d)^{-1}, \quad 0 \leq t \leq \tau = 2\pi, \end{aligned} \quad (2.27)$$

which implies

$$\psi_1(A) = 2\pi q_0(A/(2\pi)) = I_d - \frac{1}{2}A + 2 \sum_{k=1}^{\infty} \left(\frac{A}{2\pi}\right)^2 \left(\left(\frac{A}{2\pi}\right)^2 + k^2I_d\right)^{-1}. \quad (2.28)$$

Relation (2.28) is the first member of our family of rational approximations of $\psi_1(A)$.

The last may be improved by using repeatedly the same approach as above. Indeed using (2.26) and (2.22), (2.23), (2.24)

$$\begin{aligned} q_t(A) &= \frac{1}{2\pi}I + \frac{t-\pi}{2\pi}A + \sum_{i=2}^{2n+1} \frac{(2\pi)^{i-1}}{i!} B_i \left(\frac{t}{2\pi}\right) A^i + \\ &\frac{(-1)^n}{\pi} \sum_{k=1}^{\infty} \frac{A^{2n+1}(A \cos(kt) + \frac{1}{k}A^2 \sin(kt))}{k^{2n}} (A^2 + k^2I)^{-1}. \end{aligned}$$

Setting $t = 0$ we get

$$q_0(A) = \frac{1}{2\pi}I - \frac{1}{2}A + \sum_{i=2}^{2n+1} \frac{(2\pi)^{i-1}}{i!} B_i A^i + \frac{(-1)^n}{\pi} \sum_{k=1}^{\infty} \frac{A^{2n+2}}{k^{2n}} (A^2 + k^2I)^{-1}.$$

Since for $n > 1$ the odd Bernoulli numbers B_n are zeroes we have

$$\sum_{i=2}^{2n+1} \frac{(2\pi)^{i-1}}{i!} B_i A^i = \sum_{i=0}^{n-1} \frac{(2\pi)^{2i+1}}{(2(i+1))!} B_{2(i+1)} A^{2(i+1)}.$$

Hence, using $\psi_1(A) = \phi_1(A)^{-1} = 2\pi q_0(A/(2\pi))$ we arrive at the main result of the present paper

Theorem 2 *For any fixed $n > 0$ it holds*

$$\psi_1(A) = p_n(A) + 2(-1)^n \sum_{k=1}^{\infty} \left(\frac{A}{2\pi}\right)^{2(n+1)} \frac{1}{k^{2n}} \left(\left(\frac{A}{2\pi}\right)^2 + k^2I_d\right)^{-1},$$

where

$$p_n(A) = I_d - \frac{1}{2}A + \sum_{i=0}^{n-1} A^{2(i+1)} \frac{B_{2(i+1)}}{(2(i+1))!}.$$

Observe that $p_n(A)$ is the classical approximation of $\psi_1(A)$ given in Theorem 1. Also notice that the rate of convergence of the series is the same as for the series $\sum_{k=1}^{\infty} k^{-2(n+1)}$ where $2n$ is the degree of the polynomial approximation. The above result presents a rational correction of this approximation aimed to improve its convergence properties. Specifically, based on Theorem 2 we introduce the following family $\{\psi_{n,s}(A)\}_{(n,s) \in \mathbb{N} \times \mathbb{N}}$ of mixed polynomial-rational approximations of $\psi_1(A)$:

$$\psi_{n,s}(A) = p_n(A) + 2(-1)^n \left(\sum_{k=1}^s \frac{1}{k^{2n}} \left(\left(\frac{A}{2\pi} \right)^2 + k^2 I_d \right)^{-1} \right) \left(\frac{A}{2\pi} \right)^{2(n+1)}. \quad (2.29)$$

Remark 1 The above approach based on the Fourier series expansion of $q_t(z)/z$ encompasses some rational approximations of $\psi_1(z)$ which can also be derived by applying Mittag-Leffler pole decomposition (see e.g., [2] for a concise, hands-on presentation) to the function $q_0(z)$. More precisely, let us apply formula (7.54) in [2] to $q_0(z) = \psi(2\pi z) = \frac{2\pi z}{e^{2\pi z} - 1}$ with $p = 1$. The poles of our function are $\{ik\}_{k \in \mathbb{Z} \setminus \{0\}}$ and the corresponding residues are readily seen to be $\{ik\}_{k \in \mathbb{Z} \setminus \{0\}}$ as well. So we have

$$\begin{aligned} q_0(z) &= q_0(0) + zq'(0) + \sum_{k \in \mathbb{Z} \setminus \{0\}} \frac{ikz^2/(ik)^2}{z - ik} = \\ &= 1 - \pi z + \sum_{k \in \mathbb{Z} \setminus \{0\}} \frac{-iz^2}{k(z - ik)} = \\ &= 1 - \pi z + \sum_{k=1}^{\infty} \frac{2z^2}{z^2 + k^2}, \end{aligned}$$

which is exactly formula (2.13) with $t = 0$. At this point we can apply (2.12) and proceed as above (always with $t = 0$) to yield:

$$\begin{aligned} q_0(z) &= 1 - \pi z + 2 \sum_{k=1}^{\infty} z^2 \left(\sum_{i=0}^{n-1} (-1)^i \frac{z^{2i}}{k^{2i+2}} + (-1)^n \frac{z^{2n}}{k^{2n}(z^2 + k^2)} \right) = \\ &= 1 - \pi z + 2 \sum_{i=0}^{n-1} (-1)^i z^{2i+2} \sum_{k=1}^{\infty} \frac{1}{k^{2i+2}} + 2(-1)^n z^{2n} \sum_{k=1}^{\infty} \frac{1}{k^{2n}(z^2 + k^2)} \\ &= 1 - \pi z + 2 \sum_{i=0}^{n-1} (-1)^i z^{2i+2} \zeta(2i+2) + 2(-1)^n z^{2n} \sum_{k=1}^{\infty} \frac{1}{k^{2n}(z^2 + k^2)}, \quad (2.30) \end{aligned}$$

where ζ denotes the Riemann zeta function. Now recall that even-indexed Bernoulli numbers are characterized by the relation

$$B_{2\ell} = \frac{(-1)^{\ell-1} (2\ell)!}{2^{2\ell-1} \pi^{2\ell}} \zeta(2\ell) \quad (2.31)$$

(see e.g. [10], item 9.616), whereas the odd-indexed ones are zero except for $B_1 = -\frac{1}{2}$. From (2.31) we deduce

$$\zeta(2\ell) = \frac{B_{2\ell} (-1)^{\ell-1} 2^{2\ell-1} \pi^{2\ell}}{(2\ell)!} \quad (2.32)$$

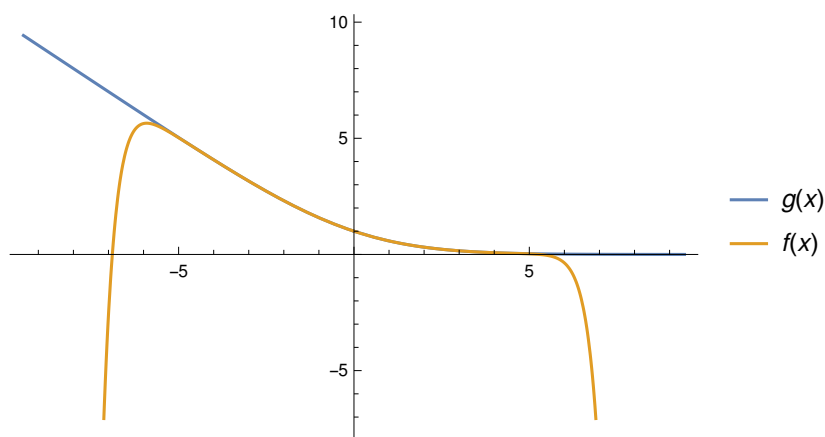


Fig. 2: Polynomial approximation $f(x) = \psi_{20,0}(x)$ against the function $g(x) = \psi_1(x)$.

labelfig1

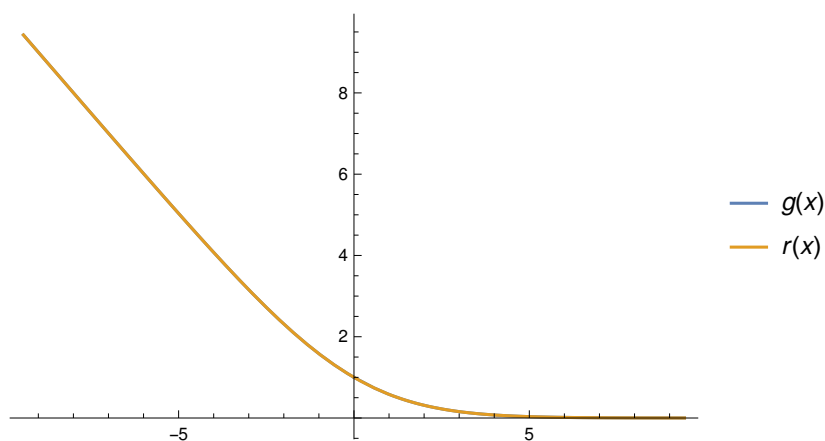


Fig. 3: Rational approximation $r(x) = \psi_{4,16}(x)$ against the function $g(x) = \psi_1(x)$. The two plots overlap.

and by plugging (2.32) with $\ell = i + 1$ in equation (2.30) we obtain

$$q_0(z) = 1 - \pi z + \sum_{i=0}^{n-1} \frac{B_{2i+2}(2\pi z)^{2i+2}}{(2i+2)!} + 2(-1)^n z^{2n} \sum_{k=1}^{\infty} \frac{1}{k^{2n}(z^2 + k^2)},$$

which is essentially the same mixed polynomial-rational development as in Theorem 2, in scalar form.

In Figures 1 and 3 we show the plot over the interval $[-3\pi, 3\pi]$ of the functions $g(x) = \psi_1(x) = \frac{x}{e^x - 1}$, its polynomial approximation $f(x) = \psi_{20,0}(x)$ and

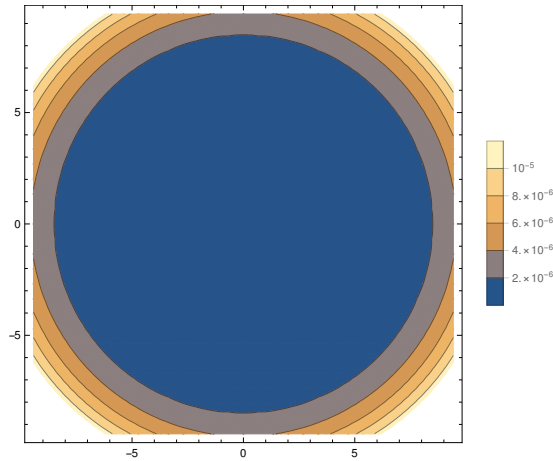


Fig. 4: Absolute error of rational approximation.

its rational approximation $r(x) = \psi_{4,16}(x)$. Clearly, the rational approximation performs better when the points are close to the border of the convergence disk of the Maclaurin series given in Theorem 1. This same phenomenon can be observed in the complex plane. In Figure 4 we illustrate the absolute error of rational approximation at complex points $x = a + ib$ with $a, b \in [-3\pi, 3\pi]$.

In the next tables we compare the accuracy of polynomial and rational approximations for computing both the matrix function $\psi_1(A)$ and the vector $\psi_1(A)\mathbf{b}$, where $A \in \mathbb{R}^{d \times d}$ is symmetric and $\mathbf{b} \in \mathbb{R}^d$. As claimed in the introduction we are interested in the case where A is structured so that we can assume that a linear system $A\mathbf{x} = \mathbf{f}$ can be solved in linear time (possibly up to logarithmic factors) with a linear storage. Our test suite is as follows:

1. A is Toeplitz tridiagonal generated as $A = \text{gallery}(\text{'tridiag'}, d, -1, 4, -1)$;
2. A is order one quasiseparable matrix generated as

$$A = 0.7 * \text{inv}(\text{gallery}(\text{'tridiag'}, d, d/2, [d : -1 : 1], d/2));$$

3. A is the Kac-Murdock-Szegö Toeplitz matrix generated as $A = ((0.8)^{\text{abs}(i-j)})$.

An accurate approximation of $\psi(A)$ is determined by computing the spectral decomposition of A . For any given $n \geq 1$ the polynomial and rational approximations of $\phi_1(A)^{-1}$ are $\psi_{n,0}(A)$ and $\psi_{m,n-m}(A)$, respectively. The corresponding normwise relative errors are

$$\text{err}_{-p_n} = \frac{\|\psi(A) - \psi_{n,0}(A)\|}{\|\psi_{n,0}(A)\|}, \quad \text{err}_{-r_{n,m}} = \frac{\|\psi(A) - \psi_{m,n-m}(A)\|}{\|\psi_{n,0}(A)\|}.$$

Tables 1, 2 and 3 show the errors evaluated using MatLab for our testing matrices. At this time we are just interested in comparing the accuracy of different approximations without incorporating fast linear solvers in our code. However, we point out that for the considered examples fast solvers exist that are expected to

| d | $err-p_{50}$ | $err-p_{50,3}$ |
|------|--------------|----------------|
| 256 | 3.33e-1 | 1.15e-12 |
| 512 | 3.33e-1 | 1.15e-12 |
| 1024 | 3.33e-1 | 1.15e-12 |
| 2048 | 3.33e-1 | 1.15e-12 |

Table 1: Errors for Example 1

| d | $err-p_{50}$ | $err-p_{50,3}$ |
|------|--------------|----------------|
| 256 | 2.25e-12 | 2.25e-12 |
| 512 | 6.42e-12 | 6.42e-12 |
| 1024 | 7.94e-3 | 2.78e-11 |
| 2048 | 2.8e77 | 2.0e-2 |

Table 2: Errors for Example 2

| d | $err-p_{50}$ | $err-p_{50,3}$ |
|------|--------------|----------------|
| 256 | 6.68e-8 | 3.71e-14 |
| 512 | 7.4e-8 | 3.78e-14 |
| 1024 | 7.6e-8 | 3.8e-14 |
| 2048 | 7.65e-8 | 3.80e-14 |

Table 3: Errors for Example 3

| γ | $err-p_{50}$ | $err-r_{50,3}$ |
|----------|--------------|----------------|
| 2 | 1.24e-11 | 1.24e-11 |
| 4 | 1.33e-10 | 1.25e-11 |
| 8 | 6.9e4 | 1.26e-11 |
| 16 | 4.15e+18 | 7.0e-11 |
| 32 | 3.1e+32 | 4.2e-9 |
| 64 | 7.0e+46 | 9.0e-7 |

Table 4: Errors for scaled companion matrices.

behave like Gaussian-elimination-based algorithms. Observe that in Example 2 for $d = 2056$ the eigenvalues are out of the disk centered at the origin of radius 2π and this explains the divergent behavior of the polynomial approximation.

In order to investigate the behavior of the different approximations under the occurrence of possibly complex eigenvalues we have compared the accuracy of polynomial and rational methods for approximating the matrix $\psi_1(A)$ where $A = \gamma F$ and F is the generator of the circulant matrix algebra, that is, the companion matrix associated with the polynomial $z^d - 1$. Since we know that the eigenvalues of F lie on the unit circle the parameter γ is used to estimate the convergence of the methods when the magnitude of eigenvalues increase. Table 4 illustrate the errors for the case $d = 1024$. The divergence of the polynomial approximation for $\gamma \geq 8$ is in accordance with the theoretical results.

Finally, for $\gamma = 64$ we consider in Table 5 the errors generated by rational approximations of increasing order. The table suggests that rational approximations

| γ | $err_r50,3$ | $err_r100,3$ | $err_r200,3$ | $err_r400,3$ |
|----------|--------------|---------------|---------------|---------------|
| 64 | 9.0e-7 | 5.7e-9 | 5.3e-11 | 1.28e-11 |

Table 5: Errors for rational approximations of increasing orders.

of higher orders are suited to give accurate results independently of the magnitude of the eigenvalues of A .

3 Bounds on the Decay of the Inverse ϕ_1 -Function

In this section we investigate the approximate rank structure of $\psi_1(A)$ for a suitable A . Specifically, as an application of Theorem 2 we can deduce *a priori* bounds on the decay of the inverse ϕ_1 -function applied to symmetric banded matrices.

Now, let $A \in \mathbb{R}^{d \times d}$ be a symmetric banded matrix. Denote as m the half-bandwidth of A , that is, $A_{i,j} = 0$ if $|i - j| > m$. It is well-known that the off-diagonal entries of $\psi_1(A)$ exhibit a decay behavior in absolute value (the same is true of any other function of A that is well-defined and sufficiently regular [3]). We can use the (n, s) -mixed polynomial-rational approximation (2.29) to give bounds on this decay behavior.

Define

$$r_{n,s}(z) = 2(-1)^n \left(\frac{z}{2\pi}\right)^{2(n+1)} \sum_{k=1}^s \frac{1}{k^{2n} \left(\left(\frac{z}{2\pi}\right)^2 + k^2\right)},$$

$$p_n(z) = 1 - \frac{1}{2}z + \sum_{i=0}^{n-1} z^{2(i+1)} \frac{B_{2(i+1)}}{(2(i+1))!}$$

which we will call the rational and the polynomial part of (2.29), respectively, and let

$$\varepsilon_{n,s}(z) = \psi_1(z) - p_n(z) - r_{n,s}(z)$$

be the (n, s) -approximation error. We have

$$|[\psi(A)]_{i,j}| \leq |[p_n(A)]_{i,j}| + |[r_{n,s}(A)]_{i,j}| + |[\varepsilon_{n,s}(A)]_{i,j}|. \quad (3.33)$$

Observe that $p_n(A)$ is a banded matrix with half-bandwidth $2nm$. So, if we choose i, j such that $|i - j| > 2nm$, then $[p_n(A)]_{i,j} = 0$ and we only need to focus on the rational and error terms.

For the rational term, let us start by giving a bound on

$$\tilde{r}_{n,s}(A) := \sum_{k=1}^s \frac{1}{k^{2n}} \left(\left(\frac{A}{2\pi}\right)^2 + k^2 I_d \right)^{-1}.$$

The matrix $A_k = \left(\frac{A}{2\pi}\right)^2 + k^2 I_d$ is positive definite with semi-bandwidth $2m$, and several exponential decay bounds for the inverse of a positive definite matrix have been proposed in the literature. Prop. 2.2 from [6], for instance, gives

$$|[(A_k)^{-1}]_{i,j}| \leq C_k \lambda_k^{|i-j|},$$

where

$$a_k = k^2, \quad b_k = \left(\frac{\rho(A)}{2\pi} \right)^2 + k^2, \quad r_k = \frac{b_k}{a_k}, \quad (3.34)$$

$$\lambda_k = \left(\frac{\sqrt{r_k} - 1}{\sqrt{r_k} + 1} \right)^{1/2m}, \quad C_k = \max \left\{ a_k^{-1}, \frac{(1 + \sqrt{r_k})^2}{2a_k r_k} \right\}, \quad (3.35)$$

where $\rho(A)$ is the spectral radius of A and $0 < a_k < b_k$ are such that the spectrum of A_k is contained in $[a_k, b_k]$. Therefore we have

$$\left| \left[\sum_{k=1}^s \frac{1}{k^{2n}} \left(\left(\frac{A}{2\pi} \right)^2 + k^2 I_d \right)^{-1} \right]_{i,j} \right| \leq \sum_{k=1}^s \frac{1}{k^{2n}} C_k \lambda_k^{|i-j|}$$

for all indices i, j . Now recall that $\left(\frac{A}{2\pi} \right)^{2(n+1)}$ is a banded matrix of bandwidth $2m(n+1)$. So we have:

$$\begin{aligned} |[r_{n,s}(A)]_{i,j}| &= \left| \sum_{\nu=1}^d [\tilde{r}_{n,s}(A)]_{i,\nu} \left[\left(\frac{A}{2\pi} \right)^{2(n+1)} \right]_{\nu,j} \right| \\ &= \left| \sum_{\nu=j-2m(n+1)}^{j+2m(n+1)} [\tilde{r}_{n,s}(A)]_{i,\nu} \left[\left(\frac{A}{2\pi} \right)^{2(n+1)} \right]_{\nu,j} \right| \\ &\leq \sum_{\nu=j-2m(n+1)}^{j+2m(n+1)} \left\| \frac{A}{2\pi} \right\|_2^{2(n+1)} \left(\sum_{k=1}^s \frac{C_k}{k^{2n}} \lambda_k^{|i-\nu|} \right) \\ &= \left\| \frac{A}{2\pi} \right\|_2^{2(n+1)} \sum_{\nu=j-2m(n+1)}^{j+2m(n+1)} \sum_{k=1}^s \frac{C_k}{k^{2n}} \lambda_k^{|i-\nu|}, \end{aligned}$$

where in the sums over ν it is understood that $1 \leq \nu \leq d$.

Let us now bound the error term. Define

$$\varepsilon_{n,s}(A) = \tilde{\varepsilon}_{n,s}(A) \left(\frac{A}{2\pi} \right)^{2(n+1)} \quad \text{where} \quad \tilde{\varepsilon}_{n,s}(A) = \sum_{k=s+1}^{\infty} \frac{1}{k^{2n}} \left(\left(\frac{A}{2\pi} \right)^2 + k^2 I_d \right)^{-1}.$$

Therefore we find that

$$|[\varepsilon_{n,s}(A)]_{i,j}| \leq \|\varepsilon_{n,s}(A)\|_2 \leq \|\tilde{\varepsilon}_{n,s}(A)\|_2 \left\| \frac{A}{2\pi} \right\|_2^{2(n+1)}.$$

Let us bound $\|\tilde{\varepsilon}_{n,s}(A)\|_2$. Let $A = UDU^H$ be the eigendecomposition of A and denote the spectrum of A as $\sigma(A)$; recall that $\sigma(A)$ is real. We have

$$\|\tilde{\varepsilon}_{n,s}(A)\|_2 \leq \|\tilde{\varepsilon}_{n,s}(D)\|_2 = \max_{x \in \sigma(A)} |\tilde{\varepsilon}_{n,s}(x)|$$

and moreover

$$|\tilde{\varepsilon}_{n,s}(x)| = \left| \sum_{k=s+1}^{\infty} \frac{1}{k^{2n}} \left(\left(\frac{x}{2\pi} \right)^2 + k^2 \right)^{-1} \right| \leq \sum_{k=s+1}^{\infty} \frac{1}{k^{2n+2}} = \zeta(2n+2) - \sum_{k=1}^s \frac{1}{k^{2n+2}},$$

from which we deduce

$$|[\varepsilon_{n,s}(A)]_{i,j}| \leq \left\| \frac{A}{2\pi} \right\|_2^{2(n+1)} \left(\zeta(2n+2) - \sum_{k=1}^s \frac{1}{k^{2n+2}} \right),$$

where $\zeta(s)$ is the Riemann zeta function.

Summing up the following estimates are obtained for the entries of $\psi_1(A)$.

Theorem 3 *Let $A \in \mathbb{R}^{d \times d}$ be a symmetric banded matrix with half-bandwidth m . For all $(n, s) \in \mathbb{N} \times \mathbb{N}$ it holds*

$$|[\psi(A)]_{i,j}| \leq |[r_{n,s}(A)]_{i,j}| + |[\varepsilon_{n,s}(A)]_{i,j}|, \quad |i-j| > 2mn,$$

where

$$|[r_{n,s}(A)]_{i,j}| \leq \left\| \frac{A}{2\pi} \right\|_2^{2(n+1)} \sum_{\nu=j-2m(n+1)}^{j+2m(n+1)} \sum_{k=1}^s \frac{C_k}{k^{2n}} \lambda_k^{|i-\nu|}$$

and

$$|[\varepsilon_{n,s}(A)]_{i,j}| \leq \left\| \frac{A}{2\pi} \right\|_2^{2(n+1)} \left(\zeta(2n+2) - \sum_{k=1}^s \frac{1}{k^{2n+2}} \right),$$

and C_k and λ_k are given in (3.34),(3.35).

To illustrate the significance of these bounds we present in Figure 5 numerical comparisons with other existing bounds deduced from [3]. Recall that these latter estimates are based on a theoretical result on the best degree- k polynomial approximation of the function $\psi_1(z)$ on $[-1, 1]$ that cannot be explicitly computed. The corresponding best polynomial approximation error satisfies

$$E_k(\psi_1) \leq \frac{2M(\chi)}{\chi^k(\chi-1)}$$

and depend on a parameter χ that defines a Bernstein ellipse in the complex plane, where the function is analytic. For the case considered in Figure 5 a good choice is $\chi = 12$. If the spectrum of the matrix is not contained in $[-1, 1]$, one needs to scale the matrix, that is, apply the function $\psi_{1,\xi}(z) = \frac{\xi z}{e^{\xi z} - 1}$ to A/ξ , for a suitable choice of ξ . Then the poles of the function closest to zero are at $\pm \frac{2\pi i}{\xi}$; the minor semi axis β of the ellipse should be chosen a little smaller than $\frac{2\pi}{\xi}$ and $\chi = \beta + \sqrt{\beta^2 + 1}$.

We see that the proposed mixed polynomial-rational approximation and the best polynomial approximation exhibit a similar decaying profile.

4 Conclusion and Future Work

In this paper we have introduced a family of rational approximations of the inverse of the ϕ_1 -function encountered in exponential integration methods. This family extends customary approximations based on the Taylor series by showing better convergence properties. Therefore, the novel formulas are particularly suited when applied for computing the inverse of the ϕ_1 matrix function of a structured matrix admitting fast and numerically robust linear solvers. Mixed polynomial-rational approximations of a meromorphic function based on the Dunford-Cauchy integral

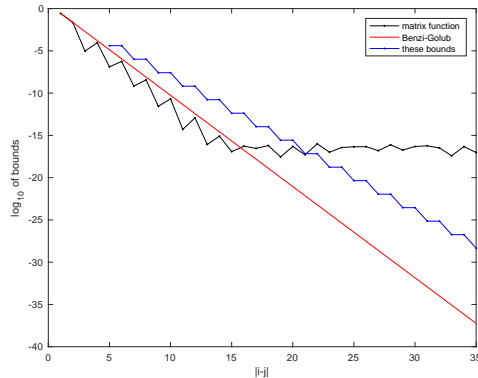


Fig. 5: Bounds for a symmetric tridiagonal matrix A with spectrum lying in $[-1, 1]$.

formula that are suited for computation with rank-structured matrices have been recently proposed in [15]. Theoretical and computational comparisons between the two families of approximations of $\psi_1(z) = \phi_1(z)^{-1}$ is an ongoing work.

Also, rational Krylov methods have been shown to be very effective for the computation of $\phi_1(A)$ especially in the case where the linear operator A comes from stiff systems of differential equations [9]. Since the features of these methods depend heavily on the properties of the associated rational function approximation problem we argue that the analysis of the proposed family of rational approximations of $1/\phi_1(z)$ can give insights for the design of novel rational Krylov schemes. In particular, the zeros of our rational approximants of $\psi_1(z)$ seem to provide some eligible sets of poles to be used in the design of parallel rational Krylov algorithms for the computation of $\phi_1(A)$. In Figure 6 we show the set of roots of $\psi_{0,80}(z)$ and $\psi_{1,80}(z)$. We see that all the zeros are symmetrically distributed in the right complex half-plane which makes the Krylov process suitable for matrices whose eigenvalues lie in the left complex half-plane [9].

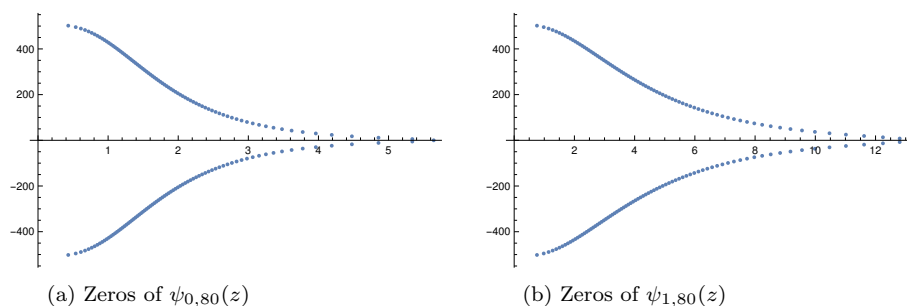
Finally, comparison of the approaches based on the Mittag-Leffler theorem and the rational Carathéodory-Fejér approximation [20] for evaluating $\psi_1(z)$ would also be interesting.

Acknowledgments. Most of the first author's work was done while at XLIM-MATHIS, Université de Limoges (UMR CNRS 7252) and on secondment in the AriC group at LIP, ENS de Lyon (CNRS, ENS Lyon, Inria, UCBL).

We thank I. V. Tikhonov for his valuable remarks.

References

1. M. Abramowitz and I. A. Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55 of *National Bureau of Standards Applied*

Fig. 6: Plots of the zeros of some rational approximants of $\psi_1(z)$

Mathematics Series. For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C., 1964.

2. G. B. Arfken and H. J. Weber. *Mathematical methods for physicists*, 1999.
3. M. Benzi and G. H. Golub. Bounds for the entries of matrix functions with applications to preconditioning. *BIT*, 39(3):417–438, 1999.
4. P. Boito, Y. Eidelman, and L. Gemignani. Efficient solution of parameter dependent quasiseparable systems and computation of meromorphic matrix functions. ArXiv:1611.09107v2. In Press in *Numerical Linear Algebra With Applications*; DOI: 10.1002/nla.2141, 2017.
5. E. Celledoni, H. Marthinsen, and B. Owren. An introduction to Lie group integrators—basics, new developments and applications. *J. Comput. Phys.*, 257(part B):1040–1061, 2014.
6. S. Demko, W. F. Moss, and P. W. Smith. Decay rates for inverses of band matrices. *Math. Comp.*, 43(168):491–499, 1984.
7. K. S. Eckhoff. Accurate reconstructions of functions of finite regularity from truncated Fourier series expansions. *Math. Comp.*, 64(210):671–690, 1995.
8. Y. Eidelman, I. Gohberg, and I. Haimovici. *Separable type representations of matrices and fast algorithms. Vol. 1*, volume 234 of *Operator Theory: Advances and Applications*. Birkhäuser/Springer, Basel, 2014. Basics. Completion problems. Multiplication and inversion algorithms.
9. T. Göckler and V. Grimm. Uniform approximation of φ -functions in exponential integrators by a rational Krylov subspace method with simple poles. *SIAM J. Matrix Anal. Appl.*, 35(4):1467–1489, 2014.
10. I. S. Gradshteyn and I. M. Ryzhik. *Table of integrals, series, and products*. Academic press, 2014.
11. N. J. Higham and L. Lin. Matrix functions: a short course. In *Matrix functions and matrix equations*, volume 19 of *Ser. Contemp. Appl. Math. CAM*, pages 1–27. Higher Ed. Press, Beijing, 2015.
12. M. Hochbruck and A. Ostermann. Exponential integrators. *Acta Numer.*, 19:209–286, 2010.
13. A. Iserles and S. P. Nørsett. On the solution of linear differential equations in Lie groups. *R. Soc. Lond. Philos. Trans. Ser. A Math. Phys. Eng. Sci.*, 357(1754):983–1019, 1999.
14. W. Magnus. On the exponential solution of differential equations for a linear operator. *Comm. Pure Appl. Math.*, 7:649–673, 1954.
15. S. Massei and L. Robol. Decay bounds for the numerical quasiseparable preservation in matrix functions. *Linear Algebra Appl.*, 516:212–242, 2017.
16. H. Munthe-Kaas. Runge-Kutta methods on Lie groups. *BIT*, 38(1):92–111, 1998.
17. H. Munthe-Kaas. High order Runge-Kutta methods on manifolds. *Appl. Numer. Math.*, 29(1):115–127, 1999.
18. A. I. Prilepko, D. G. Orlovsky, and I. A. Vasin. *Methods for solving inverse problems in mathematical physics*, volume 231 of *Monographs and Textbooks in Pure and Applied Mathematics*. Marcel Dekker, Inc., New York, 2000.
19. Eidelman Yu. S., Tikhonov I. V., and Sherstyukov V. B. Application of Bernoulli polynomials in non-classical problems of mathematical physics. In *Systems of Computer Mathematics and their Applications*, pages 223–226. Smolensk, 2017. (Russian).

-
20. T. Schmelzer and L. N. Trefethen. Evaluating matrix functions for exponential integrators via Carathéodory-Fejér approximation and contour integrals. *Electron. Trans. Numer. Anal.*, 29:1–18, 2007/08.
 21. I. V. Tikhonov. Uniqueness theorems in linear nonlocal problems for abstract differential equations. *Izv. Ross. Akad. Nauk Ser. Mat.*, 67(2):133–166, 2003.
 22. I. V. Tikhonov and Yu. S. Eidelman. An inverse problem for a differential equation in a Banach space and the distribution of zeros of an entire function of Mittag-Leffler type. *Differ. Uravn.*, 38(5):637–644, 717, 2002.