

Efficient computation of the sinc matrix function for the integration of second-order differential equations

Lidia Aceto^{1†} and Fabio Durastante^{2*†}

²Dipartimento di Scienze e Innovazione Tecnologica, University of Eastern
Piedmont, Viale T. Michel, 11, Alessandria, 15121, AL, Italy.

^{2*}Dipartimento di Matematica, University of Pisa, Via F. Buonarroti,
1/C, Pisa, 56127, PI, Italy.

*Corresponding author(s). E-mail(s): fabio.durastante@unipit.it;

Contributing authors: lidia.aceto@uniupo.it;

†These authors contributed equally to this work.

Abstract

This work deals with the numerical solution of systems of oscillatory second-order differential equations which often arise from the semi-discretization in space of partial differential equations. Since these differential equations exhibit (pronounced or highly) oscillatory behavior, standard numerical methods are known to perform poorly. Our approach consists in directly discretizing the problem by means of Gautschi-type integrators based on **sinc** matrix functions. The novelty contained here is that of using a suitable rational approximation formula for the **sinc** matrix function to apply a rational Krylov-like approximation method with suitable choices of poles. In particular, we discuss the application of the whole strategy to a finite element discretization of the wave equation.

Keywords: second-order differential equation; matrix function; sinc; rational Krylov methods

MSC2010: 65L06 , 15A16 , 41A20

1 Introduction

We consider the numerical solution of the system of multi-frequency oscillatory second-order differential equations of the form

$$\begin{cases} \mathbf{y}''(t) + A\mathbf{y}(t) = \mathbf{f}(t), \\ \mathbf{y}(t_0) = \mathbf{y}_0, \\ \mathbf{y}'(t_0) = \mathbf{y}_1, \end{cases} \quad (1)$$

where $A \in \mathbb{R}^{n \times n}$ is a symmetric and positive semi-definite matrix implicitly containing the dominant frequencies of the problem. Usually, oscillatory differential equations arise from semi-discretization in space of d -dimensional partial differential equations

$$\begin{cases} u_{tt}(\mathbf{x}, t) + \mathcal{A}(u(\mathbf{x}, t)) = f(\mathbf{x}, t), & (\mathbf{x}, t) \in \Omega \times (t_0, t_f], \Omega \subseteq \mathbb{R}^d, \\ u(\mathbf{x}, t_0) = u_0(\mathbf{x}), & \mathbf{x} \in \Omega \\ u_t(\mathbf{x}, t_0) = v_0(\mathbf{x}), & \mathbf{x} \in \Omega \\ \mathcal{B}(u(\mathbf{x}, t)) = 0, & (\mathbf{x}, t) \in \partial\Omega \times [t_0, t_f], \end{cases}$$

for $\mathcal{A}(\cdot)$ a differential operator with respect to the space variables – involving either ordinary or fractional derivatives, and $\mathcal{B}(\cdot)$ the relevant boundary conditions. The technique used to reduce partial differential equations to (large) ordinary differential equations is known as method of lines. In particular, the so-called *longitudinal method of lines* separates the problem of discretization in space from the problem of evolution in time by using the intermediate step in which one discretizes only in space but maintains continuous time. Thus, the resulting semi-discrete problem is an initial value problem of the form (1).

Several integration strategies start from the rewriting of (1) as a first-order system. By introducing the variable $\mathbf{z}(t) = [\mathbf{y}(t), \mathbf{y}'(t)]^T$ we are able to transform the second-order differential problem into the following system of first-order

$$\mathbf{z}'(t) + B\mathbf{z}(t) = \mathbf{g}(t), \quad \mathbf{z}(t_0) = \mathbf{z}_0$$

where, for O the all-zeros matrix and I the identity matrix,

$$B = \begin{bmatrix} O & -I \\ A & O \end{bmatrix}, \quad \mathbf{g}(t) = \begin{bmatrix} \mathbf{0} \\ \mathbf{f}(t) \end{bmatrix}, \quad \mathbf{z}_0 = \begin{bmatrix} \mathbf{y}_0 \\ \mathbf{y}_1 \end{bmatrix}.$$

Then, the solution can be obtained by reading the first row of the two-by-two block formula

$$\mathbf{z}(t) = \exp(-(t - t_0)B) \mathbf{z}_0 + \int_{t_0}^t \exp(-(t - \xi)B) \mathbf{g}(\xi) d\xi. \quad (2)$$

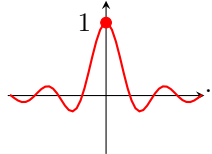
This choice is theoretically viable, and has been used many times in combination with Padé expansion for the matrix exponential, e.g., see [1]. Furthermore, by applying suitable quadrature formulas to (2) one produces the so-called exponential integrators schemes, for which we refer to the review [2]. To have efficient evaluations we require

routines for computing products of matrix functions times a vector of Krylov-type. Nevertheless, we may have efficiency difficulties. Indeed, if A is symmetric one needs to work with a matrix that is similar to a skew-symmetric matrix and, as observed in [3, 4], polynomial Krylov approaches for the matrix exponential have indeed a slower convergence rate than in the symmetric case. To overcome this drawback in [3] the proposed approach turns to polynomial Krylov methods with restart.

Our approach in this work consists in directly discretizing the problem (1) by means of Gautschi-type integrators [5, 6]; see Section 2. We focus on the numerical integration scheme given by

$$\mathbf{y}_{n+1} - 2\mathbf{y}_n + \mathbf{y}_{n-1} = h^2 \left(\operatorname{sinc} \frac{h\sqrt{A}}{2} \right)^2 (-A\mathbf{y}_n + \mathbf{f}_n),$$

with the *unnormalized sinc function* defined by

$$\operatorname{sinc} r = \begin{cases} \frac{\sin r}{r} & \text{for } r \neq 0 \\ 1 & \text{for } r = 0 \end{cases} = \text{img} \quad (3)$$


The novelty contained here is that of using a rational approximation formula for (3) to instead apply a rational Krylov-like approximation method with suitable choices of poles; see Sections 3 and 4.

Then, in Section 5 we discuss the application of the whole strategy to a finite element discretization of the wave equation. Finally, in Section 6 we perform a numerical exploration of the proposed approach.

1.1 Notation

To simplify the reading of the paper, we report some notations we adopt throughout the paper. With $\|\cdot\|$ we indicate the vector and matrix 2-norm, while with $\|\cdot\|_{\Sigma}$ we denote the function uniform norm (or sup norm) over the set Σ , i.e., $\|f\|_{\Sigma} = \sup\{|f(s)| : s \in \Sigma\}$. For a generic matrix A , we denote the field-of-values of A by

$$W(A) = \{z \in \mathbb{C} : z = \mathbf{v}^H A \mathbf{v}, \mathbf{v} \in \mathbb{C}^n, \|\mathbf{v}\| = 1\}.$$

2 Gautschi-type methods

In this section we recall the main features of time-stepping procedures that are generally known as the Gautschi-type methods [5, 6]. By making the usual position in which \mathbf{y}_n is an approximation to the value $\mathbf{y}(t_0 + nh)$, $h = (t_f - t_0)/n_t$, $n = 0, 1, \dots, n_t$, one can introduce multi-step methods of a given *trigonometric order* by looking at linear functionals $\mathfrak{L} \in \mathcal{C}^s$ of trigonometric order p , relative to period T , i.e., annihilating all

trigonometric polynomials of degree $\leq p$ with period T or, in other terms, such that

$$\mathfrak{L}1 = \mathfrak{L} \cos\left(r \frac{2\pi}{T} t\right) = \mathfrak{L} \sin\left(r \frac{2\pi}{T} t\right) = 0, \quad r = 1, 2, \dots, p.$$

In general, we can compare methods of trigonometric order p with methods having algebraic order $2p$ [5, p. 381]. Among these are the extrapolation methods of Störmer-type of trigonometric order p and with uniform step size h which take the following form

$$\mathbf{y}_{n+1} + \alpha_{p,1}(v)\mathbf{y}_n + \alpha_{p,2}(v)\mathbf{y}_{n-1} = h^2 \sum_{\ell=1}^{2p-1} \beta_{p,\ell}(v)\mathbf{y}_{n+1-\ell}'', \quad (v = 2\pi h/T), \quad (4)$$

where the expressions of the $\alpha_{p,\ell}(\cdot)$ and $\beta_{p,\ell}(\cdot)$ power series are themselves in the seminal paper of Gautschi [5, Section 5]. In some cases they can be expressed in closed form as, for example, when $p = 1$. Indeed in the latter case they are the following

$$\alpha_{1,1}(v) = -2, \quad \alpha_{1,2}(v) = 1, \quad \beta_{1,1}(v) = (\text{sinc } v/2)^2. \quad (5)$$

Therefore, for the continuous problem (1) we can specify the extrapolation Störmer scheme of trigonometric order 1 as

$$\mathbf{y}_{n+1} - 2\mathbf{y}_n + \mathbf{y}_{n-1} = h^2 \psi(h^2 A)(-A\mathbf{y}_n + \mathbf{f}_n), \quad (6)$$

for $\psi(v^2) = (\text{sinc } v/2)^2$. Since this is a two-step scheme, we can construct its equivalent one-step formulation on a staggered-grid as

$$\begin{cases} \mathbf{v}_{n+1/2} = \mathbf{v}_n + \frac{h}{2}\psi(h^2 A)(-A\mathbf{y}_n + \mathbf{f}_n), \\ \mathbf{y}_{n+1} = \mathbf{y}_n + h\mathbf{v}_{n+1/2}, \\ \mathbf{v}_{n+1} = \mathbf{v}_{n+1/2} + \frac{h}{2}\psi(h^2 A)(-A\mathbf{y}_{n+1} + \mathbf{f}_{n+1}) \end{cases} \quad \mathbf{v}_0 = \sigma(h^2 A)\mathbf{y}_1,$$

where \mathbf{v}_n represents the discretization of the velocity $\mathbf{y}'(t_0 + nh)$ and $\sigma(v^2) = \text{sinc } v$. If we concatenate the last equation coming from the previous time step with the first equation of the subsequent time step and we take as initial guess

$$\mathbf{v}_{1/2} = \sigma(h^2 A)\mathbf{y}_1 + \frac{h}{2}\psi(h^2 A)(-A\mathbf{y}_0 + \mathbf{f}_0), \quad (7)$$

we can simplify the scheme to

$$\begin{cases} \mathbf{y}_{n+1} = \mathbf{y}_n + h\mathbf{v}_{n+1/2}, & n \geq 0, \\ \mathbf{v}_{n+1/2} = \mathbf{v}_{n-1/2} + h\psi(h^2 A)(-A\mathbf{y}_n + \mathbf{f}_n), & n > 0. \end{cases} \quad (8)$$

Therefore, the computational effort that we have to make in case we want to apply the scheme (8) is the repeated computation of operations of the same type with

$$\psi(z) = (\text{sinc } \sqrt{z}/2)^2, \quad \sigma(z) = \text{sinc } \sqrt{z}. \quad (9)$$

Remark 1. For $v \rightarrow 0$, using (3) it is easy to check that the coefficient $\beta_{1,1}(v)$ in (5) tends to 1. Therefore, the corresponding method reduces to the usual Störmer-Verlet-leapfrog method

$$\mathbf{y}_{n+1} - 2\mathbf{y}_n + \mathbf{y}_{n-1} = h^2(-A\mathbf{y}_n + \mathbf{f}_n). \quad (10)$$

Arguments similar to those above give rise to the simplified scheme

$$\begin{cases} \mathbf{y}_{n+1} = \mathbf{y}_n + h\mathbf{v}_{n+1/2}, & n \geq 0 \\ \mathbf{v}_{n+1/2} = \mathbf{v}_{n-1/2} + h(-A\mathbf{y}_n + \mathbf{f}_n), & n > 0 \end{cases}$$

with initial guess

$$\mathbf{v}_{1/2} = \mathbf{y}_1 + \frac{h}{2}(-A\mathbf{y}_0 + \mathbf{f}_0).$$

This is the computationally most economic implementation, and numerically more stable than (10); see [7, p. 472].

3 Rational Krylov methods

An efficient way of performing the computations in (7)-(8) is using subspace projection methods. In the following, we denote by V_k an orthogonal matrix whose columns $\mathbf{v}_1, \dots, \mathbf{v}_k$ span an arbitrary Krylov subspace $\mathcal{W}_k(A, \mathbf{v})$ of dimension k . Then we can obtain an approximation of any $g(A; t)\mathbf{v}$ – and analogously for any of the functions in (9) – by

$$g(A; t)\mathbf{v} \approx \mathbf{g}_k = V_k g(V_k^T A V_k; t) V_k^T \mathbf{v}. \quad (11)$$

For selecting between different methods of obtaining the approximation (11) we have to make suitable choices of the projection spaces $\mathcal{W}_k(A, \mathbf{v})$. In complete generality, if we can select a set of scalars – called poles – $\{\zeta_1, \dots, \zeta_{k-1}\} \subset \mathbb{C}$ (the extended complex plane), that are not eigenvalues of A , then we can define the polynomial

$$q_{k-1}(z) = \prod_{j=1}^{k-1} (\zeta_j - z),$$

and consider as $\mathcal{W}_k(A, \mathbf{v})$ the rational Krylov subspace of order k associated with A, \mathbf{v} and q_{k-1} defined by

$$\mathcal{Q}_k(A, \mathbf{v}) = [q_{k-1}(A)]^{-1} \mathcal{K}_k(A, \mathbf{v}), \quad (12)$$

for

$$\mathcal{K}_k(A, \mathbf{v}) = \text{Span}\{\mathbf{v}, A\mathbf{v}, \dots, A^{k-1}\mathbf{v}\}$$

the standard polynomial Krylov space.

The crucial point of the entire procedure is therefore the choice of the appropriate poles for the pair function and matrix under consideration. The principle to be

guided by in the choice is that of the expression of the error committed using the approximation (11) for a given Krylov space/set of poles.

Theorem 1 (Near optimality, [8, 9]). *Let g be analytic in a neighborhood of a compact set $\Sigma \supseteq W(A)$ the field-of-value of A . Then the rational Krylov approximation (11) defined using the space (12) satisfies*

$$\|g(A; t)\mathbf{v} - \mathbf{g}_k\| \leq 2C\|\mathbf{v}\| \min_{r_k \in \mathbb{P}_{k-1}/q_{k-1}} \|g - r_k\|_{\Sigma},$$

with a constant $2 \leq C \leq 1 + \sqrt{2}$. If A is Hermitian, the result holds even with $C = 1$ and $\Sigma \supseteq \Lambda(A) \cup \Lambda(V_k^T A V_k)$, for $\Lambda(\cdot)$ the spectrum.

The goal is therefore to choose the poles so as to obtain the best rational approximation of the function g on the set Σ . For this purpose, a possible choice to obtain an upper bound is to fix a rational approximation for the function sought and to choose the poles of this approximation as poles of the method. With this in mind, in the next section we deal with obtaining these approximations in different ways.

4 Four approximations to the sinc function

The heart of the approach is therefore that of finding an approximation for the sinc function (3) to either directly approximate the matrix function-vector product or to determine the poles $\{\zeta_j\}_j$ to build the Krylov space (12) and the related approximation (11). Such results can be obtained by using the expression of the sinc function in terms of the confluent hypergeometric function (Section 4.1) or by inverting its Fourier transform (Section 4.2). While a closed form expression of the diagonal Padé approximant of the sinc function exist, see [10, p. 367], it involves the computation of determinants of matrices whose entries are binomial coefficients. This task requires symbolical manipulations and does not produce an expression of the approximation error; we give a tabulation of few of them in Appendix A.

4.1 Padé-type approximants

For the following analysis we need to introduce the confluent hypergeometric function of the first kind. This can be defined by the generalized hypergeometric series

$${}_1F_1(a; b; x) = \sum_{n=0}^{+\infty} \frac{(a)_n}{(b)_n} \frac{x^n}{n!},$$

where, as usual, $(w)_s = w(w+1) \cdots (w+s-1)$ denotes the Pochhammer symbol. It is straightforward to verify that when $b = a$ this function coincides with the exponential function, i.e.,

$$e^x = {}_1F_1(a; a; x). \quad (13)$$

Now, by virtue of the fact that

$$\operatorname{sinc} z = \frac{\sin z}{z} = \frac{e^{-iz} - e^{iz}}{-2iz}, \quad (14)$$

we may deduce a first rational approximation for the sinc function by the Padé approximants for the exponential. Similarly, this can be done by considering the Padé approximants for the ${}_1F_1(1; 2; \cdot)$. In fact, we know that if $\operatorname{Re} b > \operatorname{Re} a > 0$, the confluent hypergeometric function can be represented as an integral (see, e.g., [11, eq. (1) p. 255])

$${}_1F_1(a; b; x) = \frac{\Gamma(b)}{\Gamma(a)\Gamma(b-a)} \int_0^1 e^{xt} t^{a-1} (1-t)^{b-a-1} dt.$$

Setting $a = 1$, $b = 2$ and recalling that for every positive integer n , $\Gamma(n) = (n-1)!$, we obtain

$${}_1F_1(1; 2; x) = \int_0^1 e^{xt} dt = \frac{e^x - 1}{x}.$$

Choosing $x = -2iz$, from the previous and (14) we get

$$e^{iz} {}_1F_1(1; 2; -2iz) = e^{iz} \frac{e^{-2iz} - 1}{-2iz} = \frac{e^{-iz} - e^{iz}}{-2iz} = \operatorname{sinc} z. \quad (15)$$

Alternatively, setting $x = \pm iz$, we have

$$\frac{1}{2} \left({}_1F_1(1; 2; iz) + {}_1F_1(1; 2; -iz) \right) = \operatorname{sinc} z. \quad (16)$$

In all these three cases, rational approximations to the sinc function can be determined by using the results provided by Luke in [12] which concern the $[n/n]$ -Padé approximant of ${}_1F_1(1; b; -x)$. As for the remainder, since the confluent hypergeometric functions can be linked to the so-called φ -functions by

$${}_1F_1(1; j+1; x) = j! \varphi_j(x), \quad j = 0, 1, 2, \dots,$$

we can use the error estimate deriving from the diagonal Padé approximant to $\varphi_j(x)$ and given in [13, Lemma 2] which is sharper than the one reported in [12].

4.1.1 Padé approximants for the exponential function

As already mentioned in (13), the exponential function can be expressed in terms of a confluent hypergeometric function as follows

$$e^{-x} = {}_1F_1(1; 1; -x).$$

Therefore, from [12, eqs. (13)–(15) p. 15] we can readily obtain its $[n/n]$ -Padé approximant, i.e.,

$$e^{-x} = \frac{G_n(-x)}{G_n(x)} + S_n(x), \quad (17)$$

where

$$G_n(x) = \frac{(2n)!}{n!} {}_1F_1(-n; -2n; x) \quad (18)$$

and the remainder is given in terms of modified Bessel functions

$$S_n(x) = (-1)^{n+1} \pi e^{-x} \frac{I_{n+1/2}(x/2)}{K_{n+1/2}(x/2)} = \frac{(-1)^{n+1} n! n!}{(2n)!(2n+1)!} x^{2n+1} + \mathcal{O}(x^{2n+2}).$$

Remark 2. Now, it is worth to observe that the generalized Laguerre polynomial of degree n is defined by suitable confluent hypergeometric function as follows (see, e.g., [14, Eq. (13.6.19)])

$$L_n^{(\alpha)}(x) \triangleq \binom{n+\alpha}{n} {}_1F_1(-n; \alpha+1; x). \quad (19)$$

While for $\alpha > -1$ the $\{L_n^{(\alpha)}(x)\}_n$ are orthogonal polynomials on $[0, +\infty)$ with respect to the weight $x^\alpha e^{-x}$, for $\alpha \in \mathbb{C}$ this is no longer true, that is, they are no longer orthogonal and possess simple complex zeros; see, e.g., the discussion in [15]. This is the case of interest in the following computations.

Then, we can express (18) as

$$G_n(x) = \frac{(2n)!}{n!} \binom{-n-1}{n}^{-1} L_n^{(-2n-1)}(x).$$

Consequently, (17) becomes

$$e^{-x} = \frac{L_n^{(-2n-1)}(-x)}{L_n^{(-2n-1)}(x)} + S_n(x).$$

This relation together with (14) gives rise to our first rational approximation for the sinc function

$$\text{sinc } z = -\frac{1}{2iz} \frac{\left(L_n^{(-2n-1)}(-iz)\right)^2 - \left(L_n^{(-2n-1)}(iz)\right)^2}{L_n^{(-2n-1)}(iz)L_n^{(-2n-1)}(-iz)} + \hat{S}_n(z), \quad (20)$$

where the remainder is given by

$$\hat{S}_n(z) \triangleq \frac{S_n(iz) - S_n(-iz)}{-2iz} \approx \frac{n!n!}{(2n)!(2n+1)!} z^{2n}. \quad (21)$$

Therefore, for each value of n , the poles can be chosen as the zeros of the polynomial $L_n^{(-2n-1)}(iz)$ and its conjugate plus $z = 0$, i.e.,

$$\mathcal{E}_n = \{\zeta \in \mathbb{C} : L_n^{(-2n-1)}(i\zeta) = 0 \vee L_n^{(-2n-1)}(-i\zeta) = 0\} \cup \{0\}.$$

We compare in Figure 1 the rational approximation in (20) with the one obtained determining in a symbolic way the Padé expansion of the sinc function (see Appendix A) and observe a good agreement of the two approximations.

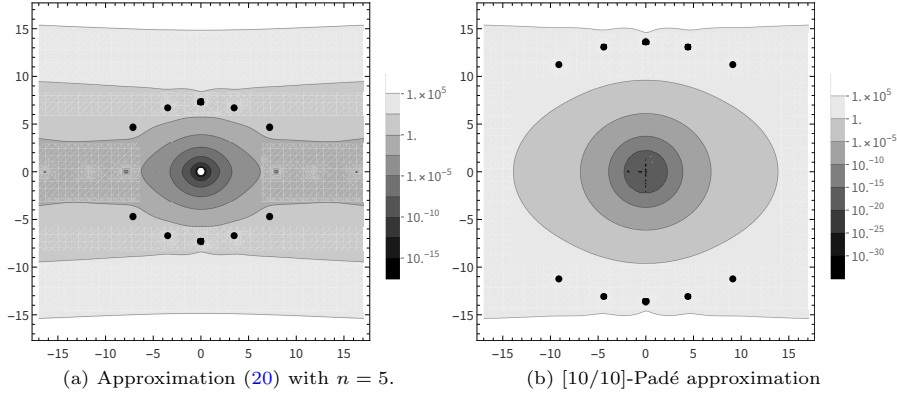


Fig. 1: Absolute error for the approximation (20) and for the $[10/10]$ -Padé approximation of the sinc function computed numerically in the $[-17, 17] \times [-17i, 17i]$ region. The color-map is \log_{10} -scale and the bold black dots represent the poles in the two cases; for the approximation (20) the pole in 0 is denoted in white to make it visible against its surrounding.

4.1.2 Padé approximants for ${}_1F_1(1; 2; \cdot)$

Following the results given by Luke in [12, Section 3] and again the higher order term that can be obtained from [13, Lemma 2], we can easily obtain the $[n/n]$ -Padé approximant to the function ${}_1F_1(1; 2; -x)$. Thus, we write

$${}_1F_1(1; 2; -x) = \frac{\mathcal{A}_n(x)}{\mathcal{B}_n(x)} + R_n(x), \quad (22)$$

where

$$\mathcal{B}_n(x) = x^n \frac{\Gamma(n+2)}{\Gamma(2n+2)} {}_2F_0(-n, n+2; ; -x^{-1}), \quad (23)$$

$$R_n(x) = (-1)^{n+1} \left(\frac{n!}{\sqrt{2}(2n+1)!} \right)^2 x^{2n+1} + (-1)^{n+1} \frac{n+1}{2n+3} \left(\frac{n!}{\sqrt{2}(2n+1)!} \right)^2 x^{2n+2} + \mathcal{O}(x^{2n+3}). \quad (24)$$

Note that we have deliberately omitted the explicit form of the polynomial $\mathcal{A}_n(x)$ as it is useless for our analysis. Since we are interested in understanding which are the zeros of the denominator of this $[n/n]$ -Padé approximant, we now focus on $\mathcal{B}_n(x)$. By virtue of the fact that (see [11, eq. (3) p. 257])

$${}_2F_0(-n, n+2; ; -x^{-1}) = x^{-n} \Psi(-n, -2n-1, x)$$

we can write

$$\mathcal{B}_n(x) = \frac{\Gamma(n+2)}{\Gamma(2n+2)} \Psi(-n, -2n-1, x).$$

Using [11, eq. (7) p. 257] we find

$$\Psi(-n, -2n-1, x) = \frac{\Gamma(2n+2)}{\Gamma(n+2)} {}_1F_1(-n; -2n-1; x).$$

Consequently,

$$\mathcal{B}_n(x) = {}_1F_1(-n; -2n-1; x).$$

Taking into account the formula (19) we have

$$\mathcal{B}_n(x) = \binom{-n-2}{n}^{-1} L_n^{(-2n-2)}(x).$$

Finally, noting that

$$\binom{-n-2}{n} = (-1)^n \binom{2n+1}{n}$$

it is immediate to get

$$\mathcal{B}_n(x) = (-1)^n \binom{2n+1}{n}^{-1} L_n^{(-2n-2)}(x).$$

Therefore, the above equation together with (22) and (24) leads to the following rational approximation (see (15))

$$\operatorname{sinc} z = (-1)^n \binom{2n+1}{n} \frac{\mathcal{A}_n(2iz)}{L_n^{(-2n-2)}(2iz)} e^{iz} + \hat{R}_n(z), \quad (25)$$

where

$$\hat{R}_n(z) \triangleq R_n(2iz) e^{iz} \approx -i \left(\frac{2^n n!}{(2n+1)!} \right)^2 z^{2n+1} e^{iz}. \quad (26)$$

Given this rational approximation, the poles can then be chosen as the zeros of the polynomials $L_n^{(-2n-2)}(2iz)$ for different values of n , i.e., the set

$$\mathcal{L}_n = \{\zeta \in \mathbb{C} : L_n^{(-2n-2)}(2i\zeta) = 0\}.$$

In Figure 2 we compare the rational approximation in (25) with the one obtained determining in a symbolic way the Padé expansion of the sinc function; we observe a good agreement between the two. Nevertheless, we discover a lack of symmetry of the poles obtained from the expansion (25), symmetry that should be inherited from the

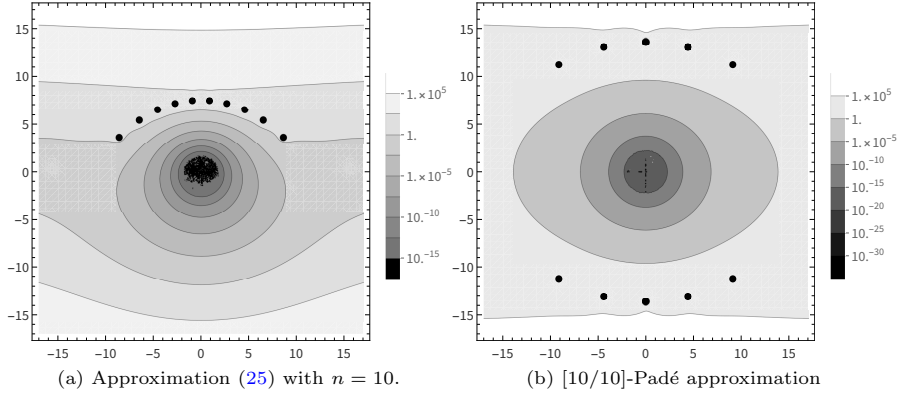


Fig. 2: Absolute error for the approximation (25) and for the [10/10]-Padé approximation of the sinc function computed numerically in the $[-17, 17] \times [-17i, 17i]$ region. The color-map is \log_{10} -scale and the bold black dots represent the poles.

parity of the sinc function. This does not happen when rewriting the sinc function as in (16). Indeed, a repeated application of (22)-(24) leads to

$$\operatorname{sinc} z = \frac{(-1)^n}{2} \binom{2n+1}{n} \frac{\mathcal{A}_n(iz)L_n^{(-2n-2)}(-iz) + \mathcal{A}_n(-iz)L_n^{(-2n-2)}(iz)}{L_n^{(-2n-2)}(iz)L_n^{(-2n-2)}(-iz)} + \tilde{R}_n(z), \quad (27)$$

with

$$\tilde{R}_n(z) = -\frac{((n+1)!)^2}{(2n+1)!(2n+3)!} z^{2n+2} + \mathcal{O}(z^{2n+3}). \quad (28)$$

This suggests considering as poles the set

$$\bar{\mathcal{L}}_n = \{\zeta \in \mathbb{C} : L_n^{(-2n-2)}(i\zeta) = 0 \vee L_n^{(-2n-2)}(-i\zeta) = 0\}.$$

A depiction of the scalar bounds for these rational approximations can be found in Appendix B.

We conclude this section by applying the previous analysis to the matrix case and summarizing it in the following result.

Proposition 2. *Let A be a symmetric and positive semi-definite matrix and $L_n^{(\alpha)}(x)$ the generalized Laguerre polynomial of degree n defined in (19). Then, the following rational approximations to the sinc matrix function can be derived:*

(i) using the Padé approximant to the exponential function and setting

$$P_n(A) := \left(L_n^{(-2n-1)}(iA) \right)^{-1} L_n^{(-2n-1)}(-iA),$$

from (20) we have

$$\operatorname{sinc}(A) \approx -\frac{1}{2i} A^{-1} [P_n(A) - (P_n(A))^{-1}] := E_n(A),$$

with the following bound for the error:

$$\|\operatorname{sinc}(A) - E_n(A)\| \leq 2 \left\| (2n+1) \left(\frac{n!}{(2n+1)!} \right)^2 z^{2n} \right\|_{\Sigma}; \quad (29)$$

(ii) using the Padé approximant for ${}_1F_1(1; 2; -2iz)$, from (25) we have

$$\operatorname{sinc}(A) \approx (-1)^n \binom{2n+1}{n} \mathcal{A}_n(2iA) \left(L_n^{(-2n-2)}(2iA) \right)^{-1} e^{iA} := F_n(A),$$

with the following bound for the error:

$$\|\operatorname{sinc}(A) - F_n(A)\| \leq 2 \left\| 2^{2n} \left(\frac{n!}{(2n+1)!} \right)^2 z^{2n+1} \right\|_{\Sigma}. \quad (30)$$

For the symmetrized version from (27) we call $\tilde{F}_n(A)$ the evaluation of the approximation in A thus obtaining

$$\|\operatorname{sinc}(A) - \tilde{F}_n(A)\| \leq 2 \left\| \frac{n+1}{4n+6} \left(\frac{n!}{(2n+1)!} \right)^2 z^{2n+2} \right\|_{\Sigma}.$$

Proof. From Theorem 1, the bounds are essentially the scalar bounds in the equations (21) for (20), (26) for (25), and (28) for (27) respectively, a part from a multiplicative factor $2C = 2$ due to A being symmetric. \square

It should be noted that, given the simplicity of computing the numerators and denominators of the rational approximations discussed here, it is also feasible to use them directly for the computation of the product $\operatorname{sinc}(A)\mathbf{v}$. This costs the same number of solutions of linear systems with shifted coefficient matrix as the case of the rational Krylov method, however with the same right-hand term and - in principle - the possibility of solving them simultaneously.

4.2 Exponential sums from the inverse Fourier transform

We recall that the Fourier transform of the $\operatorname{sinc}(r)$ function can be expressed as

$$\mathcal{F}\{\operatorname{sinc}(r)\}(k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \operatorname{sinc}(r) e^{irk} dr = \frac{1}{2} \sqrt{\frac{\pi}{2}} (\operatorname{sgn}(1-k) + \operatorname{sgn}(1+k)),$$

where

$$\operatorname{sgn}(r) = \begin{cases} +1, & r > 0, \\ 0, & r = 0, \\ -1, & r < 0, \end{cases} = \begin{array}{c} \uparrow \\ 1 \\ \circ \\ \text{---} \\ \bullet \\ \text{---} \\ \circ \\ -1 \\ \downarrow \end{array}$$

Thus we can approximate $\operatorname{sinc}(A)\mathbf{v}$ by approximating the integral of the inverse Fourier transform

$$\operatorname{sinc}(A)\mathbf{v} = \frac{1}{2} \int_{-1}^1 \exp(-ikA)\mathbf{v} dk,$$

if we then call $\{\omega_p, \ell_p\}_{p=1}^\nu$ the Gauss-Legendre quadrature points in the interval $[-1, 1]$ we can approximate it as the *exponential sum*

$$\operatorname{sinc}(A)\mathbf{v} \approx \frac{1}{2} \sum_{p=1}^\nu \omega_p \exp(-i\ell_p A)\mathbf{v}. \quad (31)$$

Another reasonable way to approximate this integral would be to use instead the Clenshaw-Curtis quadrature formula; this choice often obtains results comparable to that of Gauss-Legendre, as discussed in [16]. In the present case, numerical tests have shown us a better convergence behavior for the Gauss formula. In particular, for the latter we can show the following error bound.

Proposition 3. *The error for the approximation based on (31) for A symmetric and positive semi-definite is bounded by*

$$\left\| \operatorname{sinc}(A) - \frac{1}{2} \sum_{p=1}^\nu \omega_p \exp(-i\ell_p A) \right\| \leq \frac{\pi}{(2\nu)!} \left(\frac{\rho(A)}{2} \right)^{2\nu},$$

for $\rho(A)$ the spectral radius of A .

Proof. Since $\exp(-izk)$ has 2ν continuous derivative in $[-1, 1]$, the error of the integral approximation can be expressed as [17, Chap. 5, Page 146]

$$Q_\nu = \frac{2^{2\nu+1}(\nu!)^4}{(2\nu+1)((2\nu)!)^3} \max_{k \in [-1, 1]} \left| \frac{d^{2\nu}}{dk^{2\nu}} \exp(-ikz) \right|,$$

that is

$$Q_\nu \leq \frac{2^{2\nu+1}(\nu!)^4}{(2\nu+1)((2\nu)!)^3} |z|^{2\nu},$$

and the statement follows from

$$\lim_{\nu \rightarrow +\infty} \frac{2^{4\nu+1}(\nu!)^4}{(2\nu)!(2\nu+1)!} = \pi. \quad \square$$

To compute the sum (31) we have to compute ν matrix-exponential times vector products. The situation is analogous to the one encountered in *exponential integrators*

for first-order differential equations (2). To this end we can employ the rational approximation to the exponential function discussed in Section 4.1.1 to generate poles for computing the rational Krylov subspace $\mathcal{Q}_k(-A, \mathbf{v})$ (see again the discussion in Section 3), and then approximate the exponential sum (31) as

$$\text{sinc}(A)\mathbf{v} \approx \frac{1}{2} \sum_{p=1}^{\nu} \omega_p V_k \exp(i\ell_p A_k) V_k^T \mathbf{v}, \quad A_k = -V_k^T A V_k,$$

that can be easily adapted to the computation of (9) by discharging the computation of the square roots on the projected matrix A_k .

Remark 3. *Other viable approaches for performing the computation of the matrix-exponential vector products in (31) could be the polynomial strategy discussed in [18] that combines the usage of scaling-and-squaring techniques together with a Taylor expansion of the exponential function, or the rational expansion using the Carathéodory-Fejér approximation in [19]. Nevertheless, since we need to approximate the matrix-vector product with respect to the functions in (9) the polynomial approach would require dealing with the \sqrt{A} . On the other hand, the Carathéodory-Fejér rational approximation is designed to be uniformly optimal on negative real axis, while we need it on the imaginary one. This implies that to have sufficient accuracy, we may have to use a large number of poles to widen the region of fast convergence in the complex plane. Both alternatives are implemented in the code distributed with the paper as discussed in Section 6.*

The proposed method can be used directly for the computation of the $\sigma(z)$ function in (9) for the initial velocity in (6), on the other hand, a slightly adapted version is needed for the computation of the function $\psi(z)$ in (9) since we need to compute the matrix-vector product with respect to the square of the sinc function. This can be done again by inverting the related Fourier transform, in fact:

$$\mathcal{F} \{ \text{sinc}^2(r) \} (k) = \frac{1}{4} \sqrt{\frac{\pi}{2}} ((k-2)\text{sgn}(k-2) - 2k\text{sgn}(k) + (k+2)\text{sgn}(k+2)),$$

and thus

$$\begin{aligned} \text{sinc}^2(A)\mathbf{v} &= \frac{1}{8} \left[\int_{-2}^0 (2k+4) \exp(-ikA) dk + \int_0^2 (4-2k) \exp(-ikA) dk \right] \mathbf{v}, \\ &= \frac{1}{8} \int_{-2}^0 (2k+4) (\exp(-ikA)\mathbf{v} + \exp(ikA)\mathbf{v}) dk. \end{aligned}$$

By the same procedure, this integral can be approximated combining the Gauss-Legendre quadrature points in the interval $[-2, 0]$ with the rational Krylov method

based on the Padé poles for the exponential, thus obtaining

$$\begin{aligned} \operatorname{sinc}^2(A)\mathbf{v} \approx & \frac{1}{8}V_k \left(\sum_{p=1}^{\nu} \omega_p''(2\ell_p'' + 4) \exp(i\ell_p'' A_k)(V_k^T \mathbf{v}) \right. \\ & \left. + \sum_{p=1}^{\nu} \omega_p''(2\ell_{\nu-p+1}'' + 4) \exp(-i\ell_p'' A_k)(V_k^T \mathbf{v}) \right), \end{aligned}$$

for $A_k = -V_k^T A V_k$, and V_k a matrix whose columns span the rational Krylov subspace for A and \mathbf{v} with respect to the rational approximation to the exponential function discussed in Section 4.1.1. Note that to write the previous one we have also exploited the symmetry properties of the Gauss-Legendre weights and nodes, i.e., that the nodes and weights $\{\omega_p', \ell_p'\}_{p=1}^{\nu}$ in $[0, 2]$ are such that $\omega_p' = \omega_p''$ and $\ell_p' = -\ell_{\nu-p+1}''$, where $\{\omega_p'', \ell_p''\}_{p=1}^{\nu}$ are the Gauss-Legendre weights and nodes on the interval $[-2, 0]$.

5 A method of lines for the linear wave equation

Let us consider the acoustic wave equation

$$\begin{cases} u_{tt} - \Delta u = f, & \text{in } \Omega \times (0, T], \\ u(\mathbf{x}, 0) = u_0(\mathbf{x}), & \mathbf{x} \in \Omega, \\ u_t(\mathbf{x}, 0) = v_0(\mathbf{x}), & \mathbf{x} \in \Omega, \\ u = g_D, & \text{on } \Gamma_D \times [0, T], \\ \nabla u \cdot \hat{\mathbf{n}} = g_N, & \text{on } \Gamma_N \times [0, T], \end{cases} \quad (32)$$

with $T > 0$, $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, an open bounded domain with Lipschitz continuous boundary $\partial\Omega = \Gamma_D \cup \Gamma_N$, $\Gamma_D \cap \Gamma_N = \emptyset$, and normal vector $\hat{\mathbf{n}}$. To arrive at a system of second-order differential equations of the form (1) we apply a space semi-discretization using a Finite Element Method (FEM). Let $H \triangleq \mathbb{L}^2(\Omega)$ be the space square integrable function for which we denote the scalar product as $\langle \cdot, \cdot \rangle$, and the corresponding norm $\|\cdot\|_H$, then we denote by

$$\begin{aligned} \mathbb{H}^1(\Omega) &= \{v \in H : \partial v / \partial x_i \in H, i = 1, \dots, d\}, \\ \mathbb{H}_{\Gamma_D}^1(\Omega) &= \{v \in \mathbb{H}^1(\Omega) : v = 0 \text{ on } \Gamma_D\} \triangleq V, \end{aligned}$$

together with its dual $\mathbb{H}_{\Gamma_D}^{-1}(\Omega) \triangleq V^*$. The semi-discrete weak formulation of (32) can therefore be expressed as

$$\begin{aligned} \langle u_{tt}(\cdot, t), w \rangle + a(u(\cdot, t), w) &= \langle f(\cdot, t), w \rangle + \langle g_N, w \rangle, & \forall w \in V, \\ \langle u(\cdot, 0), w \rangle &= \langle u_0, w \rangle, & \forall w \in H, \\ \langle u_t(\cdot, 0), w \rangle &= \langle v_0, w \rangle, & \forall w \in H, \end{aligned} \quad (33)$$

with

$$a(u, w) = \int_{\Omega} \nabla u \cdot \nabla w \, d\mathbf{x}, \quad \forall u, w \in V,$$

where we use the space-time function spaces

$$\mathbb{L}^2(0, T; X) = \left\{ \phi : (0, T) \rightarrow X \text{ s.t. } \int_0^T \|\phi(t)\|_X^2 \, dt < \infty \right\},$$

in which we consider any $\phi(t) \triangleq \phi(\cdot, t)$ a function of the space variable only for fixed values of t , thus having

$$\begin{aligned} u_{tt} &\in \mathbb{L}^2(0, T; V^*), \\ u_t &\in \mathbb{L}^2(0, T; H) \cap \mathcal{C}([0, T]; V) \triangleq \mathcal{H} \cap \mathcal{C}([0, T]; V), \\ u &\in \{v \in \mathbb{L}^2(0, T; V) : v_t \in \mathcal{H}\} \cap \mathcal{C}([0, T]; V) \triangleq \mathcal{V} \cap \mathcal{C}([0, T]; V). \end{aligned}$$

We must now fix the spaces of discrete approximation to obtain (1) as a Galerkin approximation of the variational formulation (33).

5.1 FEM approximation spaces

For a set $\Omega \subseteq \mathbb{R}^2$ we consider a triangulation \mathcal{T}_h with maximum mesh edge length h of the domain Ω and the space of linear finite elements, i.e., the space of piecewise linear continuous finite element, for which we recall that the degrees of freedom (DoFs) needed to uniquely identify a function in \mathcal{V}_h are its values at the vertices, see Figure 3,

$$\mathcal{V}_h = \{v \in \mathbb{H}^1(\Omega) : \forall \kappa \in \mathcal{T}_h, v|_{\kappa} \in \mathbb{P}^1[\mathbf{x}]\} = \text{Span}\{\phi_j(\mathbf{x})\}_{j=1}^{N_{\text{DoFs}}} \subseteq \mathcal{V} \cap \mathcal{C}([0, T]; V).$$

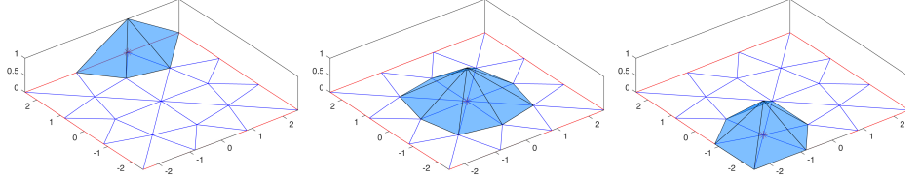


Fig. 3: Basis functions for the piecewise linear continuous finite elements.

The discrete version of (33) can then be written in \mathcal{V}_h as

$$M\mathbf{u}''(t) + K\mathbf{u}(t) = \mathbf{F}(t) + \mathbf{G}(t),$$

with

$$(M)_{i,j} = \int_{\Omega} \phi_i \phi_j \, d\mathbf{x}, \quad (K)_{i,j} = \int_{\Omega} \nabla \phi_i \cdot \nabla \phi_j \, d\mathbf{x},$$

and

$$(\mathbf{F}(t))_i = \int_{\Omega} f(\mathbf{x}, t) \phi_i \, d\mathbf{x}, \quad (\mathbf{G}(t))_i = \int_{\Omega} g_N(\mathbf{x}, t) \phi_i \, d\mathbf{x}.$$

To impose Dirichlet boundary conditions we use the so-called null-space technique, that is, we eliminate Dirichlet conditions from the problem by operating on the matrix. We build a matrix Z spanning the null-space of the columns of the matrix representing the Dirichlet condition equations, i.e., we restrict the matrices as

$$M_c = Z^T M Z, \quad K_c = Z^T K Z, \quad \mathbf{F}_c(t) = Z^T (\mathbf{F}(t) + \mathbf{G}(t) - K \mathbf{g}_D),$$

where \mathbf{g}_D is the function evaluating the Dirichlet boundary conditions at time t on the relevant degree of freedoms and taking value zero elsewhere. We can now rewrite the scheme (8) for this specific case using these matrix expressions.

6 Numerical experiments

In the following three sections we first consider the different choices of poles for computing the sinc function of a matrix and the approach based on exponential sums; Section 6.1. Then we consider a synthetic example of a second-order differential equation of the type (1) to show how error analysis allows us to choose the number of poles we need to maintain the convergence order of the method (8); Section 6.2. Finally, we consider the application of the method to the solution of the discretized finite element version of the wave equation (32); Section 6.3. The code for reproducing the experiments is available in the GitHub repository github.com/Cirdans-Home/rationalsecondorder. All the numerical experiments have been run on a Linux laptop with an Intel[®] Core[™] i7-8750H CPU at 2.20GHz with 16 Gb of RAM using Matlab R2022a.

6.1 Benchmark of the pole selection

In this preliminary section we first compare the convergence of the rational Krylov method for the choices of poles discussed in Section 4. In Table 1 we report information on test matrices we are going to use for the test. These are all discretizations of the

Table 1: Test matrices: Finite Difference discretization of the Laplacian, centered differences for the 1D case, 5-point stencil for the 2D case, and the linear Finite Element Method discussed in Section 5.1. λ_{\min} , λ_{\max} are the smallest and largest eigenvalues respectively.

A:	Matrix name	Size	$[\lambda_{\min}, \lambda_{\max}]$	Figure
1D	FD Laplacian	2048	$[2.3 \times 10^{-6}, 4]$	4a
2D	FD Laplacian	4096	$[0.0047, 7.9953]$	4b
2D	Linear FEM	1028	$[0.0253, 5.9008]$	4c

Laplacian, as our goal is to deal with the discretization of the wave equation discussed in Section 5. For all the poles choices we compute the relative error

$$\frac{\|\text{sinc}(A)\mathbf{v} - V_k \text{sinc}(V_k^T A V_k) V_k^T \mathbf{v}\|}{\|\text{sinc}(A)\mathbf{v}\|},$$

with V_k the matrix whose column spans the relevant rational Krylov subspace $\mathcal{Q}_k(A, \mathbf{v})$, and the matrix-argument $\text{sinc}(A)$ function computed in MATLAB as “`sinc(A) ← A \funm(A, @sin)`”. From Figure 4 we observe that the best results are given by poles obtained from the Padé diagonal approximant of sinc function here computed in a symbolic way; see Appendix A. Due to the parity of the function and by the fact that the diagonals approximant of odd order coincide with the previous even ones, we report only the even cases. The other satisfactory results are obtained with the poles \mathcal{E}_n given by the expansion discussed in Section 4.1.1, which are the complex zeros of the non orthogonal Laguerre polynomials. Also observe that the choice of poles as in $\overline{\mathcal{L}}_n$,

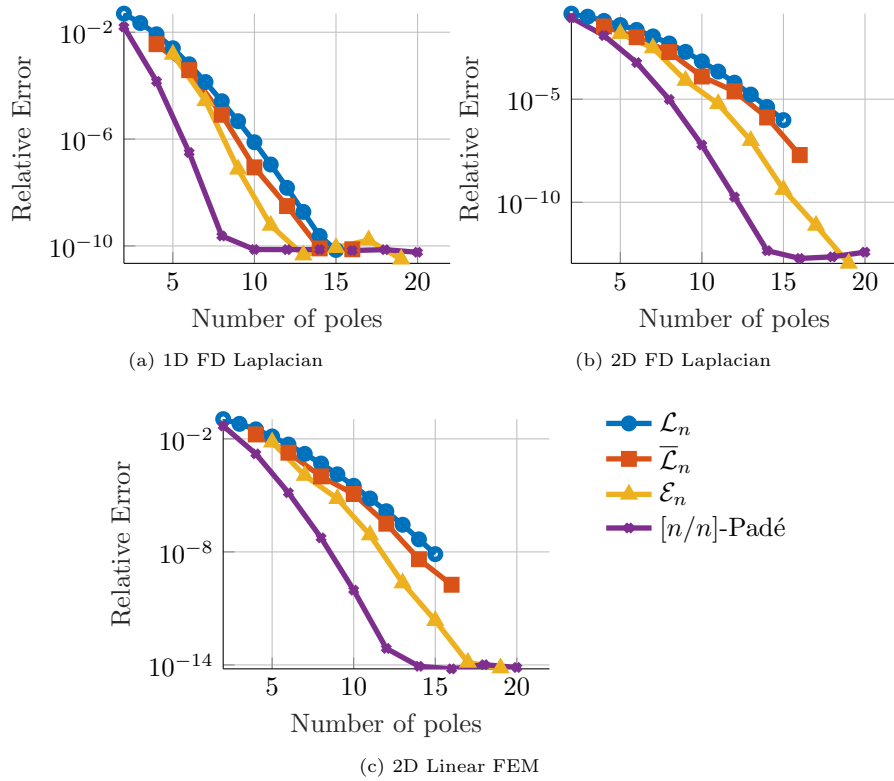


Fig. 4: Relative error for the computation of $\text{sinc}(A)\mathbf{v}$ with A the matrices described in Table 1 and \mathbf{v} a randomly generated vector. We test all the pole choices described in Section 4.

which preserve the symmetry in the approximation, slightly improves the convergence obtained using the poles of \mathcal{L}_n .

In the next set of experiments we compare the attained accuracy in terms of the relative error with the time needed to achieve it. In addition to rational Krylov-type approaches we consider also the exponential sums algorithm discussed in Section 4.2. To ensure that the strategy of approximating the matrix exponential within exponential sums does not reduce the overall accuracy we employ the cubic cost dense-matrix computation of the different exponential; this serves just as a sanity check of the overall procedure since it is indeed an expensive procedure. From the results in Figure 5 we

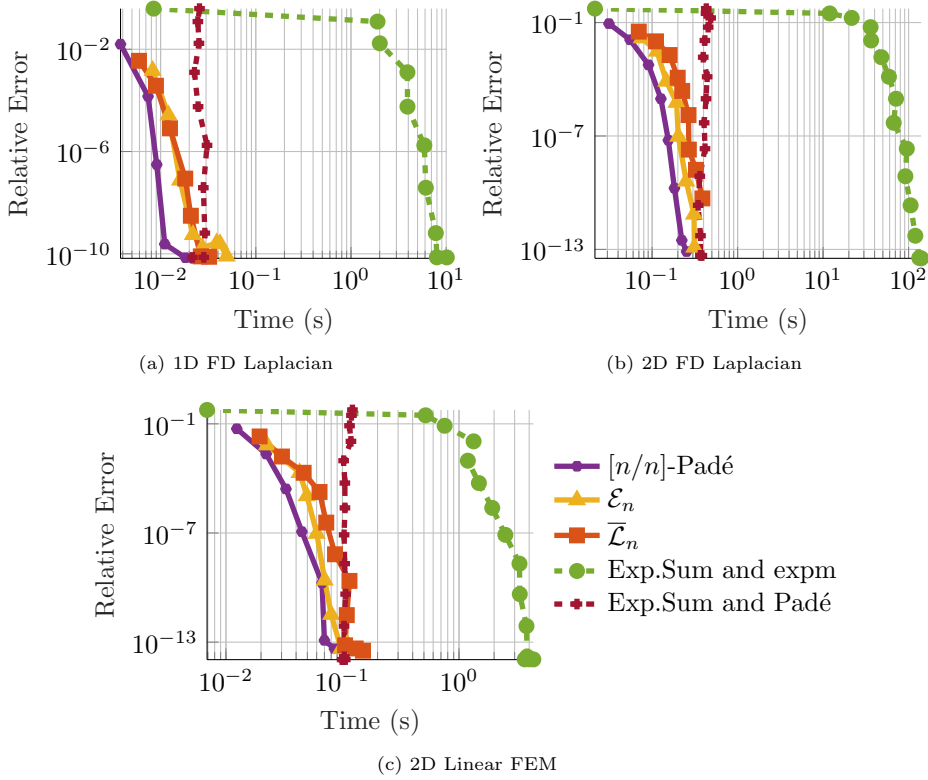


Fig. 5: Computation of the exponential sums (31) obtained using different approximations for the exponential, relative error versus time (s) graphs. To have a comparison with the results in Figure 4 we report in both pictures the results obtained with the $[n/n]$ -Padé approximation, and the rational Krylov methods with poles \mathcal{E}_n , and $\bar{\mathcal{L}}_n$. The number of poles for the $[k/k]$ -Padé approximation of the exponential is $k = 15$ for the 1D FD case and $k = 20$ for the 2D FD and Linear FEM cases. The number of Gauss-Legendre quadrature nodes goes from $\nu = 1, \dots, 15$ in all the cases.

observe that all the approximation methods reach the same accuracy for a comparable number of nodes, and that the pure rational Krylov strategies are the most cost effective.

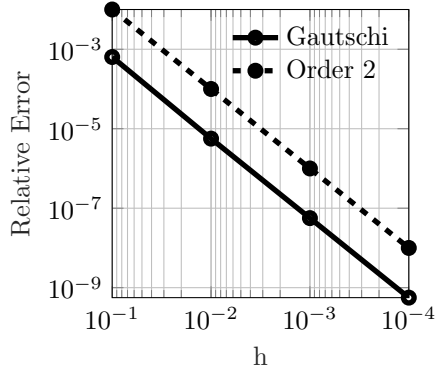
6.2 A synthetic example

In this section we consider a synthetic example of an equation of the form (1). In particular we consider as matrix $A = T_N T_N^T$ with T_N the Rutishauser matrix, i.e., the Toeplitz matrix of size N with eigenvalues in the complex plane which are approximately on the curve $2 \cos(2\theta) + 20i \sin(\theta)$, $\theta \in [0, 2\pi]$. As forcing term f we consider the function $f(t) = \frac{1}{2} \sin(t)$. The initial conditions are given by the vector identical to one for the positions and by the null vector for the velocities. To have a reference solution, we use the MATLAB integrator `ode15s` with absolute and relative tolerances equal to 10^{-12} and 3×10^{-14} , respectively. With this set of experiments firstly we want to verify that the order of Gautschi's method is preserved, secondly that it is possible to use the error analysis we have done to properly choose the number of poles. In Figure 6a we observe that using the dense-matrix computation of the matrix functions, the method (8) obtains the desired order of convergence. From Figures 6b and 6c we observe a notable consequence of the reduction of the time integration step h . If we try to obtain an increase in accuracy by reducing h , then we produce a scaling of the spectrum of the matrix A . Namely the set Σ on which compute the bounds shrinks and this makes the computation of the involved matrix function easier. In other words, we need a lower number of poles to achieve a higher precision via the reduction of the time step h . On the other hand, when we want to apply the method based on exponential sums in Figure 6d, we have a fixed cost per step which is given by the generation of the rational Krylov space for the approximation of the exponential matrix-vector products. The accuracy with which we make this product limits the final accuracy of the quadrature formula, so to maintain the second order we need a number of poles, and therefore of linear system solutions, higher than the other two cases in which the errors combine more favorably. For all strategies the error analysis allows to maintain the order 2 of the method.

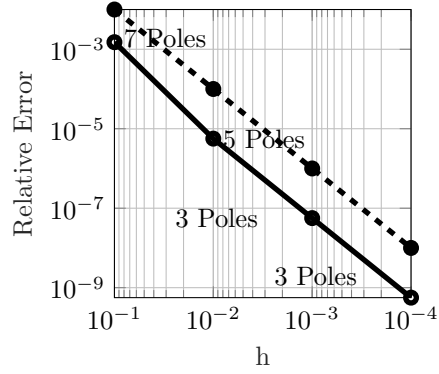
To have a comparison, we consider the exponential integrator Adams-Bashforth-Nørsett scheme of stiff order 2 from [20]. This method uses a direct computation of the involved φ functions, thus we compare it against our implementation that also uses the direct computation. To enhance the difference with respect to the cost, i.e., having to compute a matrix function of a matrix of double the size, we consider the same test problem but with $N = 100$. From the results in Table 2 we observe that for a comparable error, having to compute smaller-dimensional matrix functions has the advantage in terms of expected time.

6.3 Solution of the linear wave equation

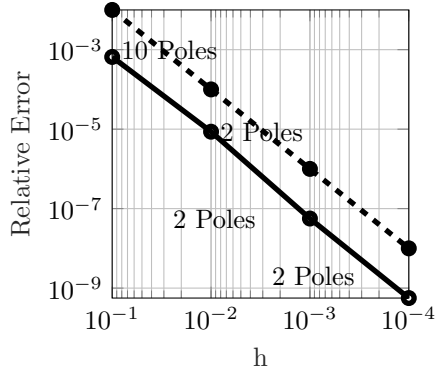
Let us now test the different strategies implemented for the computation of the sinc function to compute the matrix-vector products using the matrix functions from (9) in the scheme (8). We consider the wave equation (32) on the domain and triangular mesh depicted in the first two panels of Figure 7. We construct a problem in which the boundary conditions are homogeneous Dirichlet conditions on the whole boundary, i.e.,



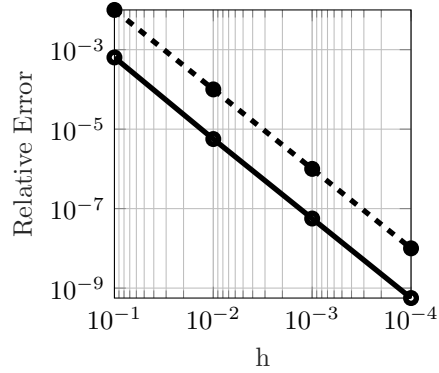
(a) Dense-matrix function computation.



(b) Poles obtained from the rational approximation (20).



(c) Poles obtained from the rational approximation (27).



(d) Exponential sum algorithm with 10 poles for the rational Krylov approximation of the exponential, and 10 quadrature points to generate the exponential sum.

Fig. 6: Relative 2-norm error at final time $T = 1$, for matrix size $N = 20$. We use the bounds from Proposition 2 to determine the number of poles we need to achieve a given tolerance. The $\|\cdot\|_{\Sigma}$ norm is estimated by evaluating the $\|\cdot\|_{\infty}$ bound on an equally spaced grid of the interval $[0, h^2 \lambda_{\max}(A)] = [0, h^2 1.2138e+03]$.

h	Gautschi		Adams-Bashforth-Nørsett	
	T (s)	Rel. Err.	T (s)	Rel. Err.
1.0e-01	4.44e-02	1.12e-04	5.49e-01	3.76e-05
1.0e-02	3.04e-02	1.12e-06	2.66e-01	4.69e-07
1.0e-03	1.42e-02	4.71e-09	2.85e-01	4.80e-09
1.0e-04	4.39e-01	6.18e-10	1.58e+00	2.35e-10

Table 2: Comparison in terms of elapsed time and two-norm relative error of the Gautschi integrator and the Adams-Bashforth-Nørsett scheme of stiff order 2 from [20].

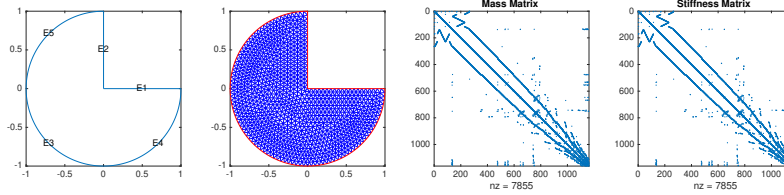


Fig. 7: From left to right, we have the domain, the triangular mesh for the solution of the wave equation (32), the mass and stiffness matrix for the \mathbb{P}^1 -element on the associated triangulation.

$\Gamma_N = \emptyset$. The initial data for the positions $\mathbf{x} = (x, y)$ is given by a perturbation of the shape

$$u_0(x, y) = 0.8 \exp(-(x+0.3)^2/0.06 - (y+0.3)^2/0.06),$$

and a zero initial velocity, i.e., $v_0(x, y) = 0$. Since we simply want to test the robustness of the routines for computing the different matrix functions, we also set the forcing term equal to zero. From the errors reported in Figure 8 it can be observed that also in

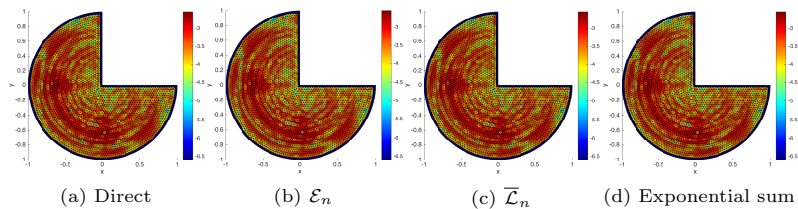


Fig. 8: Pointwise absolute error (\log_{10} -scale) computed against the solution generated with MATLAB's `solvepde` command at $T = 1$ with $h = 10^{-2}$ with respect to time, and linear FEM with largest element size of $h_{\max} = 0.03$ in the `generateMesh` program. We used four poles both for the \mathcal{E}_n and \mathcal{L}_n . The exponential sum case has been computed with 6 poles for the rational approximation of the exponential and 12 terms in the sum.

this case all the methods are capable of obtaining the same result as the direct method for a suitable (and limited) choice of the number of poles.

7 Conclusion

We have analyzed several strategies for the computation of matrix functions that appear in Gautschi-type trigonometric integrators for second-order differential equations. The analysis includes bounds to determine the number of poles of the rational approaches needed to achieve a specified accuracy. Future developments of the techniques discussed here concern the use of algorithms for the approximate solution of linear systems within rational Krylov spaces; and the possible application of the rational-Krylov methods with respect to the poles we have discussed here directly to exponential integrators for first-order systems. Furthermore, having observed the correlation between convergence

rate of the method for the matrix function, amplitude of the time discretization step and global error, one should investigate the adaptive choice of the integration step and the relative adaptive choice of the number of poles to be used in the underlying rational approximation. Along the same lines, the possibility of also considering a space adaptive FEM discretization could further improve the compromise between accuracy and execution speed of the method.

Acknowledgments. The authors are members of the INdAM research group GNCS.

Declarations

Funding. This work was partially supported by the “GNCS Research Project - INdAM” with code CUP_E53C22001930001, and by the Spoke 1 “FutureHPC & BigData” of the Italian Research Center on High-Performance Computing, Big Data and Quantum Computing (ICSC) funded by MUR Missione 4 Componente 2 Investimento 1.4: Potenziamento strutture di ricerca e creazione di “campioni nazionali di R&S (M4C2-19)” - Next Generation EU (NGEU).

Code availability. The code used to produce the results is available at <https://github.com/Cirdans-Home/rationalsecondorder>.

Appendix A Denominators of Padé diagonal approximations

Table A1 shows the denominators of the Padé expansions of the $\text{sinc}(r)$ function for even degrees from 2 to 10 calculated in closed form with Mathematica (v. 12.2.0) and the PadeApproximant function:

`Table[CoefficientList[Denominator[PadeApproximant[Sinc[x], {x, 0, n}]], x], {n, 2, 10, 2}]`

Table A1: Coefficients in the monomial basis (ascending order) of the denominators of the diagonal Padé approximations for the $\text{sinc}(r)$ function for the degree $n = 2m$, $m = 1, \dots, 5$.

n	$[n/n]$ -Padé denominators coefficients
2	$\{ 1, 0, \frac{1}{20} \}$
4	$\{ 1, 0, \frac{13}{396}, 0, \frac{5}{11088} \}$
6	$\{ 1, 0, \frac{1671}{69212}, 0, \frac{97}{351384}, 0, \frac{2623}{1644477120} \}$
8	$\{ 1, 0, \frac{2290747}{120289892}, 0, \frac{1281433}{7217393520}, 0, \frac{560401}{562956694560}, 0, \frac{1029037}{346781323848960} \}$
10	$\{ 1, 0, \frac{34046903537}{2167379498676}, 0, \frac{1679739379}{13726736824948}, 0, \frac{101555058991}{168015258737363520}, 0, \frac{3924840709}{2016183104848362240}, 0, \frac{37291724011}{11008359752472057830400} \}$

We stress that there exist a closed form expression of the diagonal Padé approximation of the sinc function that can be found in [10, p. 367] but involves computing determinants of matrices whose entries are binomial coefficients.

Computing the poles, i.e., the zeros, of such polynomials can be a delicate task. As the degree increases, the difference in absolute value of the largest and smallest coefficient drops below the machine precision. Where in general there are ad-hoc algorithms to deal with the computation of the zeros of polynomials that are difficult to represent, we resorted here in exploiting the symbolic functionalities of Mathematica to obtain a tabulation to machine precision of the poles up to degree 20. The values are available in the Git repository; see Section 6.

Appendix B Depiction of the scalar bounds

We report here a depiction of the scalar bounds for the approximations (20) and (27) on the real line. As shown in the Figure B1 there is an excellent correspondence

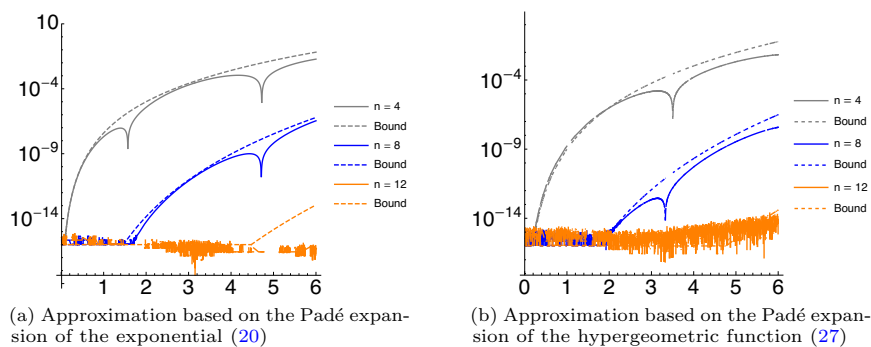


Fig. B1: Comparison of the theoretical bounds and the computed errors on the interval $[0, 6]$ for the two rational approximation based on the Padé approximation of the exponential (B1a) and the hypergeometric function (B1b) discussed respectively in Sections 4.1.1 and 4.1.2.

between the theoretical bounds (21) and (26) and the absolute error calculated with 16 significant figures. Furthermore, the results are also in good agreement with the behavior depicted in Figure 4 with respect to the choice of poles for a given matrix A . Specifically, they confirm the behavior according to which the poles obtained by the exponential expansion return a better convergence than those obtained by the expansion based on the hypergeometric function.

References

- [1] Baker, G.A., Bramble, J.H., Thomée, V.: Single step Galerkin approximations for parabolic problems. *Math. Comp.* **31**(140), 818–847 (1977) <https://doi.org/10.2307/2006116>

- [2] Hochbruck, M., Ostermann, A.: Exponential integrators. *Acta Numer.* **19**, 209–286 (2010) <https://doi.org/10.1017/S0962492910000048>
- [3] Botchev, M.A., Knizhnerman, L.A., Schweitzer, M.: Krylov subspace residual and restarting for certain second order differential equations. *arXiv* (2022). <https://doi.org/10.48550/ARXIV.2206.06909> . <https://arxiv.org/abs/2206.06909>
- [4] Hochbruck, M., Lubich, C.: On Krylov subspace approximations to the matrix exponential operator. *SIAM J. Numer. Anal.* **34**(5), 1911–1925 (1997) <https://doi.org/10.1137/S0036142995280572>
- [5] Gautschi, W.: Numerical integration of ordinary differential equations based on trigonometric polynomials. *Numer. Math.* **3**, 381–397 (1961) <https://doi.org/10.1007/BF01386037>
- [6] Hochbruck, M., Lubich, C.: A Gautschi-type method for oscillatory second-order differential equations. *Numer. Math.* **83**(3), 403–426 (1999) <https://doi.org/10.1007/s002110050456>
- [7] Hairer, E., Nørsett, S.P., Wanner, G.: *Solving Ordinary Differential Equations. I: Nonstiff Problems*, 2nd edn. Springer Series in Computational Mathematics, vol. 8, p. 528. Springer, Berlin, Heidelberg (1993)
- [8] Güttel, S.: Rational Krylov approximation of matrix functions: numerical methods and optimal pole selection. *GAMM-Mitt.* **36**(1), 8–31 (2013) <https://doi.org/10.1002/gamm.201310002>
- [9] Crouzeix, M., Palencia, C.: The numerical range is a $(1 + \sqrt{2})$ -spectral set. *SIAM J. Matrix Anal. Appl.* **38**(2), 649–655 (2017) <https://doi.org/10.1137/17M1116672>
- [10] Magnus, A., Wynn, J.: On the Padé table of $\cos z$. *Proc. Amer. Math. Soc.* **47**, 361–367 (1975) <https://doi.org/10.2307/2039747>
- [11] Erdélyi, A., Magnus, W., Oberhettinger, F., Tricomi, F.G.: *Higher Transcendental Functions. Vol. I*, 1st edn., p. 302. McGraw-Hill Book Co., Inc., New York-Toronto-London (1953). Based, in part, on notes left by Harry Bateman. <https://resolver.caltech.edu/CaltechAUTHORS:20140123-104529738>
- [12] Luke, Y.L.: *Algorithms for Rational Approximations for a Confluent Hypergeometric Function II*. Interim rept. ADA032910, Missouri University, Kansas City, Department of Mathematics, <https://apps.dtic.mil/sti/citations/ADA032910> (September 1976)
- [13] Skaflestad, B., Wright, W.M.: The scaling and modified squaring method for matrix functions related to the exponential. *Appl. Numer. Math.* **59**(3-4), 783–799 (2009) <https://doi.org/10.1016/j.apnum.2008.03.035>

- [14] *NIST Digital Library of Mathematical Functions*. <http://dlmf.nist.gov/>, Release 1.1.3 of 2021-09-15. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds. (2021). <http://dlmf.nist.gov/>
- [15] Martínez-Finkelshtein, A., Martínez-González, P., Orive, R.: On asymptotic zero distribution of Laguerre and generalized Bessel polynomials with varying parameters. In: Proceedings of the Fifth International Symposium on Orthogonal Polynomials, Special Functions and Their Applications (Patras, 1999), vol. 133, pp. 477–487 (2001). [https://doi.org/10.1016/S0377-0427\(00\)00654-3](https://doi.org/10.1016/S0377-0427(00)00654-3) . [https://doi.org/10.1016/S0377-0427\(00\)00654-3](https://doi.org/10.1016/S0377-0427(00)00654-3)
- [16] Trefethen, L.N.: Is Gauss quadrature better than Clenshaw-Curtis? *SIAM Rev.* **50**(1), 67–87 (2008) <https://doi.org/10.1137/060659831>
- [17] Kahaner, D., Moler, C., Nash, S.: *Numerical Methods and Software*. Prentice-Hall, Inc., USA (1989)
- [18] Al-Mohy, A.H., Higham, N.J.: Computing the action of the matrix exponential, with an application to exponential integrators. *SIAM J. Sci. Comput.* **33**(2), 488–511 (2011) <https://doi.org/10.1137/100788860>
- [19] Schmelzer, T., Trefethen, L.N.: Evaluating matrix functions for exponential integrators via Carathéodory-Fejér approximation and contour integrals. *Electron. Trans. Numer. Anal.* **29**, 1–18 (2007/08)
- [20] Berland, H., Skaflestad, B., Wright, W.M.: EXPINT—A MATLAB Package for Exponential Integrators. *ACM Trans. Math. Softw.* **33**(1), 4 (2007) <https://doi.org/10.1145/1206040.1206044>