NUMERICAL LINEAR ALGEBRA WITH APPLICATIONS
Numer. Linear Algebra Appl. 2003; **10**:247–270 (DOI: 10.1002/nla.287)

Solving linear initial value problems by Faber polynomials

P. Novati*

Dipartimento di Scienze Matematiche, Universit di Trieste, Via Valerio 12/1, 34100 Trieste, Italy

SUMMARY

In this paper we use the theory of Faber polynomials for solving N-dimensional linear initial value problems. In particular, we use Faber polynomials to approximate the evolution operator creating the so-called exponential integrators. We also provide a consistence and convergence analysis. Some tests where we compare our methods with some Krylov exponential integrators are finally shown. Copyright © 2002 John Wiley & Sons, Ltd.

1. INTRODUCTION

Given a matrix $A \in \mathbf{R}^{N\times N}$ and a continuous function $g:[0,T]\to\mathbf{R}^N$, we consider the linear initial value problem (IVP)

$$\begin{aligned} y'(t) &= -Ay(t) + g(t), \quad t\in[0,T] \\ y(0) &= y_0 \end{aligned} \tag{1}$$

For the remainder of the discussion we assume A to be time independent. As well known, in this situation the solution of (1) is given by

$$y(t) = \exp(-tA)y_0 + \int_0^t \exp((s-t)A)g(s)\,\mathrm{d}s \tag{2}$$

Solving (1) with a classical method involves at each step an attempt to approximate the exponential function. In particular with explicit schemes the approximation is of polynomial type, whereas implicit schemes involve a rational approximation. The drawback we want to overcome regards the fact that, generally, these approximations do not take into account the location in the complex plane of the spectrum of A, that we denote by $\sigma(A)$. This constitutes a drawback regarding especially explicit methods, because if $\sigma(A)$ is contained in a region of the complex plane where the approximation of the exponential function is not good, the method may lead to poor results, in the sense that generally a drastic reduction of the time step is required, especially in stiff cases. On the other hand, rational approximations

* Correspondence to: P. Novati, Dipartimento di Scienze Matematiche, Universit di Trieste, Via Valerio 12/1, 34100 Trieste, Italy.

Published online 30 October 2002

Received 28 February 2001
Revised 21 January 2002

arising from the use of implicit schemes generally allow to attain acceptable approximations of the matrix exponential but they require the solution of one or more linear systems at each step, that constitutes a computational disadvantage, at least unless optimal preconditioners are available.

To overcome these problems, in recent years some authors have proposed one-step integration techniques based on the direct computation of the matrix exponential operator at each step using Krylov subspace methods [1–6], that generally cost less than the solution of a linear system. Such methods are usually called *Krylov exponential integrators* (see Reference [4]). The computation is based on the projection of the matrix exponential onto the Krylov subspaces using the Arnoldi or Lanczos algorithms. Concerning the rate of convergence, these methods show a very appreciable behaviour (see e.g. Reference [2]) but they also present the typical disadvantages of the projective schemes, that is, the construction of the projection subspaces. In particular, using the classical Arnoldi algorithm to build the Krylov subspaces there is the well-known problem of the growth of the computational cost (see References [1, 6]). On the other hand, using the Lanczos algorithm there is the possibility of breakdown, with consequent failure of the method, and in general there are stability problems due to the fact that oblique projections instead of orthogonal ones are used.

Like Krylov exponential integrators, the methods for (1) we are going to introduce are one-step methods based on the computation of the matrix exponential operator, but this evaluation is performed by means of truncating Faber series defined on a certain compact subset Ω of the complex plane containing (or, more generally, approximating) $\sigma(A)$ (see References [7, 8]). As pointed out in the works just mentioned this technique is quite efficient both by convergence and computational point of views, because of the properties of approximations of Faber polynomials and by the fact that there exists a recursion they satisfy (see e.g. References [9–12]). We call *Faber exponential integrators* such kind of procedures. A similar approach, based on Chebyshev polynomials and series, that is of particular interest when A is symmetric or skew-symmetric, has recently been used in References [13–15].

The standard approaches for the solution of (1) by means of the direct computation of the matrix exponential operator are based on the use of a quadrature formula for the integral in (2). In this way, more than one matrix exponential usually must be computed at each time step. Therefore, it is fundamental to employ an efficient but not too expensive method for matrix functions. In this sense, the aim of this paper is to show the effectiveness of the Faber approximation technique as a tool for (1).

In order to clarify the notation used throughout the paper, unless otherwise specified the vector norm is always the Euclidean norm, and the matrix norm is always the spectral norm. Moreover, for a complex valued function $h : K \subseteq \mathbf{C} \to \mathbf{C}$, we define $\|h\|_K : \sup_{z \in K} |h(z)|$.

The paper is structured as follows: In Section 2 we give an outline about the theory of Faber polynomials and series. In Section 3 we consider the computation of the matrix exponential and of the matrix function $(I - \exp(-A))A^{-1}$ (the last one can arise when solving (1) with constant forcing term) by truncated Faber series. An error analysis of this kind of approximation is furnished in Section 4. In Section 5 we describe the common approaches used for the solution of (1) with a polynomial type method. Section 6 is devoted to the properties of consistency and convergence of Faber integrators. In Section 7 we provide some further numerical issues about Faber coefficients, introduced in Section 2. Finally, in Section 8 we test our method on problems arising from the discretization of a parabolic equation.

2. BACKGROUND ON FABER POLYNOMIALS AND SERIES

Let

$$\mathbf{M} := \{\Omega \subset \mathbf{C} : \Omega \text{ is compact, } \bar{\mathbf{C}}\backslash\Omega \text{ is simply connected and } \Omega \text{ contains more than one point}\}$$

Given $\Omega \in \mathbf{M}$, by the Riemann Mapping Theorem we can consider the conformal surjection

$$\psi : \bar{\mathbf{C}}\backslash\{w : |w| \leqslant \gamma\} \to \bar{\mathbf{C}}\backslash\Omega, \quad \psi(\infty) = \infty, \quad \psi'(\infty) = 1 \tag{3}$$

where the constant γ is the capacity of Ω. Let $\phi : \bar{\mathbf{C}}\backslash\Omega \to \bar{\mathbf{C}}\backslash\{w : |w| \leqslant \gamma\}$ be the inverse mapping of ψ.

The jth *Faber polynomial* is defined as the polynomial part of the Laurent expansion at ∞ of $[\phi(z)]^j$ (cf. [12], Section 2)

$$[\phi(z)]^j = z^j + \sum_{k=-\infty}^{j-1} \beta_{j,k} z^k, \quad j \geqslant 0$$

that is

$$F_j(z) := z^j + \sum_{k=0}^{j-1} \beta_{j,k} z^k, \quad j \geqslant 0.$$

As well known, in the particular case that Ω coincides with the closure of the internal part of an ellipse or with a bounded interval in the complex plane, Faber polynomials reduce to scaled and translated Chebyshev polynomials. We refer to References [16, 17] for a detailed description of these cases.

Let Γ be the boundary of Ω and, for $R > \gamma$, let $\Gamma(R)$ be the equipotential curve

$$\Gamma(R) := \{z : |\phi(z)| = R\}$$

Moreover, let us denote by $\Omega(R)$ the closure of the interior of $\Gamma(R)$, and by $\overset{\circ}{\Omega}(R)$ its internal part. For $R = \gamma$ we define $\Omega(\gamma) := \Omega$ and $\Gamma(\gamma) := \Gamma$. Let f be a function analytic on Ω. By Reference [12] (Theorem 1, p. 167), f can be uniquely expanded into a series of Faber polynomials

$$f(z) = \sum_{j=0}^{\infty} a_j(f) F_j(z), \quad z \in \Omega \tag{4}$$

where the coefficients $a_j(f)$ are called *Faber coefficients* with respect to f and the compact Ω; they are defined as

$$a_j(f) := \frac{1}{2\pi \mathrm{i}} \int_{|w|=R} \frac{f(\psi(w))}{w^{j+1}} \, \mathrm{d}w, \qquad j \geqslant 0, \quad \gamma < R \tag{5}$$

Now consider the sequence of polynomials $\{q_{m-1}(z)\}_{m \geqslant 1}$ obtained by truncating the series (4), that is

$$q_{m-1}(z) := \sum_{j=0}^{m-1} a_j(f) F_j(z) \tag{6}$$

Given a general square matrix $B \in \mathbf{C}^{N\times N}$ and a vector $u \in \mathbf{C}^N$, under the hypothesis that $\sigma(B) \subset \Omega$, it is known that the sequence

$$y_m := q_{m-1}(B)u \tag{7}$$

converges to $f(B)u$ (see Reference [7]). Moreover, by the properties of Faber polynomials, it is known that the sequence $\{q_{m-1}(z)\}_{m\geqslant 1}$ approximates asymptotically f on Ω as well as the sequence of best uniform approximation polynomials. In this sense the method (7) is said to be *asymptotically optimal* with respect to f and Ω (see e.g. Reference [16]). We call (7) *Faber series method* (FSM). Defining

$$\eta := \max\{r : r > \gamma, f \text{ analytic on } \overset{\circ}{\Omega}(r)\} \tag{8}$$

we have that a sufficient condition for the convergence of the FSM is $\sigma(B) \subset \overset{\circ}{\Omega}(\eta)$, because $q_{m-1}(z)$ converges to $f(z)$ uniformly in each compact subset contained in $\overset{\circ}{\Omega}(\eta)$ (cf. References [7, 8]). Moreover, setting the error

$$e_m = y_m - f(B)u$$

if $\sigma(B) \subset \Omega(r)$, with $r < \eta$, by Reference [7] we know that

$$\overline{\lim_{m\to\infty}} \|e_m\|^{1/m} \leqslant \frac{r}{\eta} \tag{9}$$

so that η gives a measure of the rate of convergence of the method.

3. THE COMPUTATION OF $\exp(-\delta A)v$ AND $(I - \exp(-\delta A))A^{-1}v$

Now suppose to know a certain $\Omega \in \mathbf{M}$ with capacity γ such that $\sigma(A) \subset \Omega$. By the conformal mapping theory, it is well known that if ψ is the conformal surjection relative to Ω (cf. (3)) then ψ has a Laurent expansion of the type

$$\psi(w) = w + \alpha_0 + \frac{\alpha_1}{w} + \frac{\alpha_2}{w^2} + \cdots, \quad |w| > \gamma \tag{10}$$

For the sequence $\{F_m\}_{m\geqslant 0}$ of Faber polynomials with respect to Ω, the following well-known recursion holds

$$\begin{aligned} &F_0(z) = 1, \quad F_1(z) = z - \alpha_0 \quad \text{and for } m \geqslant 2 \\ &F_m(z) = (z - \alpha_0)F_{m-1}(z) - (\alpha_1 F_{m-2}(z) \\ &\qquad + \cdots + \alpha_{m-1}F_0(z)) - (m-1)\alpha_{m-1} \end{aligned} \tag{11}$$

Now, consider first the computation of the matrix exponential. By Section 2, working with the function $f(z) = \exp(-\delta z)$, $\delta > 0$, the FSM for the computation of $\exp(-\delta A)v$ is based on the expansion

$$\exp(-\delta z) = \sum_{j=0}^{\infty} a_j(\delta)F_j(z), \quad z \in \Omega \tag{12}$$

where F_j is the jth Faber polynomial with respect to Ω and

$$a_j(\delta) := \frac{1}{2\pi \mathrm{i}} \int_{|w|=R} \frac{\exp(-\delta\psi(w))}{w^{j+1}}\, \mathrm{d}w, \qquad j \geqslant 0, \quad \gamma < R \tag{13}$$

Thus, the FSM has the form

$$w_m(\delta) = p_{m-1,\delta}(\delta A)v \quad \text{where } p_{m-1,\delta}(\delta z) := \sum_{j=0}^{m-1} a_j(\delta) F_j(z) \tag{14}$$

In order to understand the notation used, we must point out that $p_{m-1,\delta}$ is a polynomial whose coefficients depend on δ in a non-polynomial form, see (13).

Regarding the computation of the approximations $w_m(\delta)$ of (14), the following result can be easily proved by direct computation using (11) and (14).

Proposition 3.1
For $\delta > 0$ the approximations $w_m(\delta)$ can be carried out recursively by

$$\begin{aligned} &w_0(\delta) = 0, \quad w_1(\delta) = a_0(\delta)v, \quad w_2(\delta) = w_1(\delta) + \frac{a_1(\delta)}{a_0(\delta)}(A - \alpha_0 I)d_{0,\delta} \\ &w_m(\delta) = w_{m-1}(\delta) + \frac{a_{m-1}(\delta)}{a_{m-2}(\delta)}(A - \alpha_0 I)d_{m-2,\delta} - \frac{a_{m-1}(\delta)}{a_{m-3}(\delta)}\alpha_1 d_{m-3,\delta} \\ &\qquad - \cdots - \frac{a_{m-1}(\delta)}{a_0(\delta)}(m-1)\alpha_{m-2} d_{0,\delta} \quad m \geqslant 3 \end{aligned} \tag{15}$$

where $d_{0,\delta} := w_1(\delta)$, $d_{k,\delta} := w_{k+1}(\delta) - w_k(\delta)$, $(k \geqslant 1)$.

Concerning the convergence, since the exponential function is analytic on the whole complex plane, η (8) can be chosen arbitrarily large. Hence, the series (14) converges uniformly on every compact subset of $\mathbf{C}$. As consequence the corresponding method (15) converges no matter where $\sigma(A)$ is located with respect to Ω, for each $\delta > 0$. Moreover, by (9), the rate of convergence is superlinear.

Now consider the computation of $(I - \exp(-\delta A))A^{-1}v$. Since for the above description we are able to compute the matrix exponential, we can proceed in two phases computing first $u = A^{-1}v$ and then $(I - \exp(-\delta A))u$. However, this kind of approach requires the solution of a linear system which, as well known, could present some problems. It is more convenient to deal directly with the operator $\varphi(\delta A) := (I - \exp(-\delta A))(\delta A)^{-1}$. In fact, solving a linear system with a polynomial method involves the approximation of the function $1/z$, singular at 0, whereas the function φ has only a removable singularity in 0 (see Reference [8]). Hence, $\eta(\varphi)$ can be chosen arbitrarily large as for the exponential case.

Following what stated for the exponential case, we write

$$\varphi(\delta z) = \sum_{j=0}^{\infty} \bar{a}_j(\delta) F_j(z), \quad z \in \Omega \tag{16}$$

where

$$\bar{a}_j(\delta) = \frac{1}{2\pi \mathrm{i}} \int_{|w|=R} \frac{1 - \exp(-\delta\psi(w))}{\delta\psi(w) w^{j+1}} \, \mathrm{d}w, \qquad j \geqslant 0, \quad \gamma < R \tag{17}$$

Therefore, the FSM for the computation of $(I - \exp(-\delta A))A^{-1}v$ gives

$$x_m(\delta) = \delta \bar{p}_{m-1,\delta}(\delta A)v \quad \text{where } \bar{p}_{m-1,\delta}(\delta z) := \sum_{j=0}^{\infty} \bar{a}_j(\delta) F_j(z) \tag{18}$$

The approximations $x_m(\delta)$ can be clearly carried out using the recursion stated in Proposition 3.1.

4. ERROR ANALYSIS

In this section we want to provide upper bounds for the errors of (14) and (18). Since we are interested in the general case of A not diagonalizable, we work in terms of the *field of values* of A, defined as

$$F(A) := \left\{ \frac{z^H A z}{z^H z} : z \in \mathbf{C}/\{0\} \right\}$$

Moreover, as extensively explained in Reference [18], in order to analyse convergence of a matrix iterative process it is suitable to work with the field of values of the matrix involved. In fact, this tool allows to estimate not only the asymptotic performance of the iterative scheme (such informations can be derived from the spectral properties of the matrix) but it is also useful to understand the behaviour of the method for a finite number of iteration.

For the following result see Reference [19, Theorem 4.1].

Lemma 4.1
Let $d(z, F(A))$ be the distance between a point z and $F(A)$. Then

$$\|(zI - A)^{-1}\| \leqslant 1/d(z, F(A))$$

Theorem 4.1
Let Ω be convex and assume that $F(A) \subseteq \Omega(s)$, for some $\gamma \leqslant s$. Let moreover f be a function such that $s < \eta$, with η defined by (8). Given a general polynomial method $p_{m-1}(A)v \approx f(A)v$ such that $\lim_{m\to\infty} \| p_{m-1} - f \|_\Omega = 0$, for any $s < r < \eta$, the error $e_m = p_{m-1}(A)v - f(A)v$ is such that

$$\|e_m\| \leqslant \|v\| \| f - p_{m-1} \|_{\Omega(r)} \frac{r+s}{r-s} \tag{19}$$

Proof
By the definition of e_m, for any $s < r < \eta$ it is

$$e_m = \frac{1}{2\pi \mathrm{i}} \int_{\Gamma(r)} (f(z) - p_{m-1}(z))(zI - A)^{-1} v \, \mathrm{d}z$$

Then we get

$$\|e_m\| \leqslant \frac{\|f - p_{m-1}\|_{\Omega(r)}}{2\pi} \int_{|w|=r} |\psi'(w)| \|(\psi(w)I - A)^{-1} v\| \, \mathrm{d}w$$

Hence, by Lemma 4.1, we obtain

$$\|e_m\| \leqslant \frac{\|v\| \|f - p_{m-1}\|_{\Omega(r)}}{2\pi} \int_{|w|=r} \left| \frac{\psi'(w)}{\psi(w) - u(w)} \right| \mathrm{d}w$$

where $u(w) \in \Omega(s)$. Since

$$\int_{|w|=r} \left| \frac{\psi'(w)}{\psi(w) - u(w)} \right| \mathrm{d}w = \frac{1}{r} \int_{|w|=r} \left| \frac{w\psi'(w)}{\psi(w) - u(w)} \right| \mathrm{d}w$$

using the relation ([12] p. 16)

$$\left| \frac{w\psi'(w)}{\psi(w) - u(w)} \right| \leqslant \frac{r+s}{r-s}$$

we get the statement. □

Now, consider the error of approximation of the FSM. Let $\Omega \in \mathbf{M}$ with capacity γ; from now on we always assume that ϕ can be continuously extended to Γ (this holds for instance when Γ is a Jordan curve). Let $V(\Omega)$ be the *total boundary rotation* of Ω, defined as

$$V(\Omega) := \int_0^{2\pi} |d_\theta \arg(\psi(\gamma \mathrm{e}^{\mathrm{i}\theta}) - \psi(\gamma \mathrm{e}^{\mathrm{i}\theta_0}))|, \quad 0 \leqslant \theta_0 < 2\pi \tag{20}$$

We assume $V(\Omega) < +\infty$. If Ω is convex then $V(\Omega) = 2\pi$.

Remark 4.1
In what follows we often use the quantity $V(\Omega(R))$, $R > \gamma$, instead of $V(\Omega)$. However, it is known that the hypothesis $V(\Omega) < +\infty$ implies the inequality $V(\Omega(R)) < +\infty$. Indeed, as shown in Reference [20], for $R > \gamma$ sufficiently large, $\Omega(R)$ is convex, and in any case, given $\gamma < R_1 < R_2$, we have $V(\Omega(R_2)) \leqslant V(\Omega(R_1)) \leqslant V(\Omega)$.

Let f be analytic on Ω. Let us denote with $\mathbf{F}_{m-1}(f)$ the truncated Faber series (6). We have the following general result.

Proposition 4.1
Let $R < \eta$, with $\eta = \eta(f)$ defined by (8). Then, for every $\gamma \leqslant r < R$,

$$\|f - \mathbf{F}_{m-1}(f)\|_{\Omega(r)} \leqslant \frac{V}{\pi} \|f\|_{\Gamma(R)} \frac{(r/R)^m}{1 - r/R} \tag{21}$$

where $V = V(\Omega(r))$.

Proof
The bound (21) is easily derived using the well-known relations (see e.g. Reference [10])

$$\max_{z \in \Omega(r)} |F_j(z)| \leqslant \frac{V}{\pi} r^j, \quad |a_j(f)| \leqslant \frac{\|f\|_{\Gamma(R)}}{R^j} \qquad \square \tag{22}$$

Lemma 4.2
If Ω is symmetric with respect to the real axis, then, given $R > \gamma$, for the mapping ψ we have

$$\psi(-R) \geqslant \psi(-\gamma) - R + \frac{\gamma^2}{R} \tag{23}$$

Proof
Writing

$$\psi(-\gamma) = \psi(-R) + \int_{-R}^{-\gamma} \psi'(t)\,\mathrm{d}t$$

where the integral path is the real line segment $[-R, -\gamma]$, using the bound

$$|\psi'(w)| \leqslant 1 + \left(\frac{\gamma}{|w|}\right)^2, \quad |w| > \gamma \tag{24}$$

(see Reference [11]), we have

$$\begin{aligned}\psi(-\gamma) - \psi(-R) = |\psi(-\gamma) - \psi(-R)| &\leqslant \int_{-R}^{-\gamma} |\psi'(t)|\,\mathrm{d}t \\ &\leqslant \int_{-R}^{-\gamma} \left(1 + \left(\frac{\gamma}{t}\right)^2\right) \mathrm{d}t \\ &= R - \frac{\gamma^2}{R} \qquad \square\end{aligned}$$

Now, let

$$\bar{\mathbf{M}} := \left\{ \begin{array}{l} \Omega \in \mathbf{M} : \Omega \text{ is symmetric with respect to the} \\ \text{real axis, convex, and } \Gamma \text{ is a Jordan curve} \end{array} \right\} \tag{25}$$

Theorem 4.2
Let $\Omega \in \bar{\mathbf{M}}$. Assume that $F(A) \subseteq \Omega(s)$, for some $s \geqslant \gamma$. Then, for the error $e_m(\delta) = w_m(\delta) - \exp(-\delta A)v$ of (14) we have

$$\|e_m(\delta)\| \leqslant C \exp(\delta E) \left(\frac{s \exp(\delta)}{m}\right)^{m-1}, \quad m \geqslant 4s \tag{26}$$

where

$$C = 8\|v\| es \left(1 + \frac{1}{8s}\right), \quad E = 1 - \psi(-\gamma) \tag{27}$$

Proof
By (19) and (21), for $\gamma \leqslant s < r < R$ we get (using also $V/\pi = 2$ because Ω is convex)

$$\|e_m(\delta)\| \leqslant 2\|v\| \frac{(r/R)^m}{1 - (r/R)} \frac{r+s}{r-s} \max_{z \in \Gamma(R)} |\exp(-\delta z)| \tag{28}$$

If in (28) we put $r = s(1 + 1/m)$, $m \geqslant 1$, we obtain

$$s < r \leqslant 2s \quad \text{and} \quad \frac{r+s}{r-s} = 2m + 1$$

Moreover we have

$$\left(\frac{r}{R}\right)^m = \left(\frac{s}{R}\right)^m \left(1 + \frac{1}{m}\right)^m \leqslant e \left(\frac{s}{R}\right)^m \tag{29}$$

Since the exponential function is analytic in the whole complex plane, in (28) we can choose R arbitrarily large. Hence, we can put $R = m$, so that, for $m \geqslant 4s$,

$$\frac{1}{1 - r/R} \leqslant \frac{1}{1 - 2s/R} \leqslant 2 \tag{30}$$

$$2m + 1 \leqslant 2m \left(1 + \frac{1}{2m}\right) \leqslant 2m \left(1 + \frac{1}{8s}\right) \tag{31}$$

Substituting (29), (30), (31) in (28) we find

$$\|e_m\| \leqslant 8\|v\| e \left(1 + \frac{1}{8s}\right) m \left(\frac{s}{m}\right)^m \max_{z \in \Gamma(m)} |\exp(-\delta z)|, \quad m \geqslant 4s \tag{32}$$

Moreover, since Ω is convex, the same is true for each $\Omega(m)$, $m \geqslant 1$ (see Reference [20]). Hence, by the nature of the exponential function we easily get

$$\max_{z \in \Gamma(m)} |\exp(-\delta z)| = \exp(-\delta \psi(-m)) \tag{33}$$

Now, by Lemma 4.2

$$\psi(-m) \geqslant \psi(-\gamma) - m$$

and thus

$$\exp(-\delta \psi(-m)) \leqslant (\exp(\delta))^{m-1} \exp(\delta(1 - \psi(-\gamma))) \tag{34}$$

By (32), using (33) and (34) we easily get the thesis. □

Theorem 4.3
Under the hypothesis of the previous theorem, for the error $\bar{e}_m(\delta) = x_m(\delta) - (I - \exp(-\delta A))A^{-1}v$ of method (18) we have the bound

$$\|\bar{e}_m(\delta)\| \leqslant C\delta \exp(\delta E) \left(\frac{s \exp(\delta)}{m}\right)^{m-1}, \quad m \geqslant \max(4s, \bar{m}) \tag{35}$$

where $\bar{m}$ is the smallest integer such that $\psi(-\bar{m}) \leqslant 0$, C and E are defined by (27).

Proof
Since φ is analytic in the whole complex plane except for a removable singularity in 0, we can proceed as in the previous proof getting

$$\|\bar{e}_m(\delta)\| \leqslant C\delta \left(\frac{s}{m}\right)^{m-1} \max_{z \in \Gamma(m)} \left|\frac{1 - \exp(-\delta z)}{\delta z}\right|, \quad m \geqslant 4s$$

where C is defined by (27). Now, by the nature of the exponential function we get

$$\begin{aligned} \max_{z \in \Gamma(m)} \left|\frac{1 - \exp(-\delta z)}{\delta z}\right| &\leqslant \max_{z \in \Gamma(m)} \left|\frac{1 - \exp(-\delta \operatorname{real}(z))}{\delta \operatorname{real}(z)}\right| \\ &\leqslant \frac{1 - \exp(-\delta\psi(-m))}{\delta\psi(-m)} \end{aligned} \tag{36}$$

and thus, for each $m \geqslant \bar{m}$,

$$\frac{1 - \exp(-\delta\psi(-m))}{\delta\psi(-m)} \leqslant \exp(-\delta\psi(-m))$$

Using Lemma 4.2 as before we get the thesis.

5. THE SOLUTION OF THE SYSTEM

Now let us see how to use the methods described in Section 3 in order to solve (1). By (2), we can express $y(t + \delta)$ as

$$y(t + \delta) = \exp(-\delta A)y(t) + \int_0^\delta \exp(-(\delta - \tau)A)g(t + \tau)\,\mathrm{d}\tau \tag{37}$$

Formula (37) can be used as the basis for a time-stepping procedure.

5.1. First approach

The standard approach for the numerical implementation of (37) consists of using a general quadrature formula of the type

$$\int_0^\delta \exp(-(\delta - \tau)A)g(t + \tau)\,\mathrm{d}\tau \approx \delta \sum_{j=1}^{p} \mu_j \exp(-(\delta - \tau_j)A)g(t + \tau_j) \tag{38}$$

where the τ_j and μ_j are the quadrature nodes and weights, respectively, in $[0, \delta]$. Except the case of the trapezoidal rule, the use of any other higher-order quadrature formula requires the evaluation of more than one matrix exponential at each step. For this reason, some Krylov exponential integrators consider g as a constant on the interval of integration or use approximate projection formulas. This is intended to avoid the construction of more than one Krylov subspaces sequence (see e.g. Reference [1]).

Using the FSM to compute the matrix exponentials of (38), if y_n is an approximation of $y(t)$, fixed a certain integer $m \geqslant 1$ we consider the following one-step method for the approximation of $y(t+\delta)$:

$$y_{n+1} := p_{m-1,\delta}(\delta A) y_n + \delta \sum_{j=1}^{p} \mu_j \, p_{m-1,\delta-\tau_j}((\delta - \tau_j)A) g(t+\tau_j) \tag{39}$$

where the polynomial $p_{m-1,\delta}$ is defined by (14). We call *Faber exponential integrator* the method (39) and we denote it by F$[m,k]$, where k indicates that the error of the quadrature formula is of the type $O(\delta^k)$.

Remark 5.1
Clearly, depending on the quadrature rule, in order to get an error of the type $O(\delta^k)$, the function g must be smooth enough. Hence, whenever we refer to the method F$[m,k]$, we always assume that g satisfies the necessary smoothness properties.

Although formula (39) could appear quite complicated and expensive, we can use the recursion stated in Proposition 3.1 to carry out it, so that the total computation requires generally $(p+1)(m-1)$ matrix–vector products at each step. Note that if in (38) the point δ is a quadrature node then (39) requires only $p(m-1)$ matrix by vector products.

5.2. Second approach

If in the integral term of (37) we consider constant the function g in $[t, t+\delta]$, i.e. we approximate $g(t+\tau)$ with $g(t)$ for $\tau \in [0, \delta]$, we get

$$y(t+\delta) \approx \exp(-\delta A) y(t) + \left(\int_0^{\delta} \exp(-(\delta - \tau)A) \, \mathrm{d}\tau \right) g(t)$$

Hence, from the identity

$$\int_0^{\delta} \exp(-(\delta - \tau)A) \, \mathrm{d}\tau = A^{-1}(I - \exp(-\delta A)) = \delta \varphi(\delta A) \tag{40}$$

we can use the relation

$$y(t+\delta) \approx \exp(-\delta A) y(t) + \delta \varphi(\delta A) g(t)$$

as the basis for the following one-step integration scheme

$$y_{n+1} := p_{m-1,\delta}(\delta A) y_n + \delta \, \bar{p}_{m-1,\delta}(\delta A) g(t_n) \tag{41}$$

where the polynomial $\bar{p}_{m-1,\delta}$ is defined by (18). Obviously (41) works well only if $g(t_n)$ approximates well $g(t)$ for $t \in [t_n, t_{n+1}]$, so that the time step δ does not need to be drastically

reduced. We call F[m] the method (41). To carry out (41) we can use the recursion of Proposition 3.1. This allows to achieve y_{n+1} with $2(m-1)$ matrix vector applications at each step.

In the case of time-constant forcing $g(t)=g$, by (37) and (40) we get

$$\begin{aligned} y(t+\delta) &= \exp(-\delta A)y(t) + A^{-1}(I-\exp(-\delta A))g \\ &= \exp(-\delta A)y(t) + \delta\varphi(\delta A)g \end{aligned} \tag{42}$$

so that it is natural to use the method F[m] described by (41), obviously with $g(t_n)=g$ for each n, i.e.

$$y_{n+1} := p_{m-1,\delta}(\delta A)y_n + \delta\bar{p}_{m-1,\delta}(\delta A)g \tag{43}$$

For the particular case of $g=0$, the method (43) becomes simply

$$y_{n+1} := p_{m-1,\delta}(\delta A)y_n \tag{44}$$

which obviously requires $m-1$ matrix by vector applications at each step. It is interesting to observe that in this case F[m] generalizes the explicit Runge–Kutta method, in the sense that if $\psi(w)=w$ then $p_{m-1,\delta}$ is the truncated Taylor expansion of the exponential function in a neighbourhood of 0.

6. CONSISTENCY AND CONVERGENCE

Before studying the consistency properties of the two approaches, we must give error bounds for the approximations of the exponential and the function φ, as $\delta\to 0$.

Proposition 6.1
Let $\Omega\in\bar{\mathbf{M}}$. For the FSM, as $\delta\to 0$ we have

$$\|e_m(\delta)\| = O(\delta^m), \quad \|\bar{e}_m(\delta)\| = O(\delta^{m+1}) \tag{45}$$

Proof
By (28),

$$\|e_m(\delta)\| \leqslant 2\|v\|\exp(-\delta\psi(-R))\,\frac{(r/R)^m}{1-r/R}\,\frac{r+s}{r-s}, \quad s<r<R<+\infty$$

Defining first $r=2s$, we get

$$\frac{r+s}{r-s}=3$$

so that

$$\|e_m(\delta)\| \leqslant \text{const}\,\exp(-\delta\psi(-R))\,\frac{(r/R)^m}{1-r/R}, \quad r<R<+\infty$$

Since we can choose arbitrarily $R > r$, for $\delta < 1$ let us define $R = r/\delta$. In this way, as $\delta \to 0$

$$-\delta\psi(-R) = 2s + O(\delta) \tag{46}$$

so that

$$\|e_m(\delta)\| = O(\delta^m)$$

that proves the first part of (45).

Proceeding as before, by (19), (21) and (36), defining $r = 2s$ and $R = r/\delta$, we find

$$\|\bar{e}_m(\delta)\| \leqslant \text{const}\, \frac{1 - \exp(-\delta\psi(-r/\delta))}{\delta\psi(-r/\delta)}\, \frac{\delta^{m+1}}{1-\delta}$$

Using (46) we get

$$\frac{1 - \exp(-\delta\psi(-r/\delta))}{\delta\psi(-r/\delta)} = O(1)$$

that completes the proof. □

Every one-step method can be written in the form

$$y_{n+1} = y_n + \delta\Phi(t_n, y_n; \delta)$$

and the local discretization error is defined as

$$d(t,\delta) = \frac{y(t+\delta) - y(t)}{\delta} - \Phi(t, y(t); \delta), \quad t \in [0, T]$$

We have the following results.

Theorem 6.1
The F$[m,k]$ method is consistent with the problem (1) with consistency order equal to $q = \min(m,k) - 1$.

Proof
For the F$[m,k]$ method we have

$$\Phi(t_n, y_n; \delta) := \frac{p_{m-1,\delta}(\delta A) - I}{\delta} y_n + \sum_{j=1}^{p} \mu_j\, p_{m-1,\delta-\tau_j}((\delta - \tau_j)A) g(t_n + \tau_j) \tag{47}$$

Thus by (37) we have

$$\begin{aligned} d(t,\delta) &= \frac{(\exp(-\delta A) - p_{m-1}(\delta A))}{\delta}\, y(t) \\ &\quad + \frac{1}{\delta}\int_0^{\delta} \exp(-(\delta - \tau)A) g(t+\tau)\, \mathrm{d}\tau - \textstyle\sum_{j=1}^{p} \mu_j\, p_{m-1,\delta-\tau_j}((\delta - \tau_j)A) g(t + \tau_j) \\ &= d_1(t,\delta) + d_2(t,\delta) + d_3(t,\delta) \end{aligned}$$

where

$$d_1(t,\delta) := \frac{\exp(-\delta A) - p_{m-1}(\delta A)}{\delta} y(t)$$

$$d_2(t,\delta) := \frac{1}{\delta}\int_0^{\delta} \exp(-(\delta-\tau)A) g(t+\tau)\,\mathrm{d}\tau - \sum_{j=1}^{p} \mu_j \exp(-(\delta-\tau_j)A) g(t+\tau_j)$$

$$d_3(t,\delta) := \sum_{j=1}^{p} \mu_j \exp(-(\delta-\tau_j)A) g(t+\tau_j) - \sum_{j=1}^{p} \mu_j p_{m-1,\delta-\tau_j}((\delta-\tau_j)A) g(t+\tau_j)$$

A bound for $d_1(t,\delta)$ can be obtained using (45), that is

$$\|d_1(t,\delta)\| \leqslant k_1 \delta^{m-1} \|y\|_{C[0,T]} \tag{48}$$

where $\|y\|_{C[0,T]} := \max_{t\in[0,T]} \|y(t)\|$. For $d_2(t,\delta)$, by hypothesis

$$\|d_2(t,\delta)\| \leqslant k_2 \delta^{k-1} \tag{49}$$

For $d_3(t,\delta)$, proceeding as for $d_1(t,\delta)$, we find

$$\|d_3(t,\delta)\| \leqslant \left(\sum_{j=1}^{p} |\mu_j|\right) k_1 \delta^{m-1} \|g\|_{C[0,T]} \tag{50}$$

Finally, by (48), (49), (50), it follows:

$$\max_{t\in[0,T]} \|d(t,\delta)\| \leqslant O(\delta^q) \qquad \square$$

Example 6.1
If we use the Faber iteration with $m=5$ together with the Simpson rule for (38), which implies $k=5$, we obtain a method of order 4 which requires 12 matrix by vector applications at each step.

Theorem 6.2
The F[m] method is consistent with the problem (1). If $g:[0,T]\to\mathbf{R}^N$ is of class C^1, the consistency order is equal to 1. If $g(t)=g$ is constant, then the consistency order is $m-1$.

Proof
For the F[m] method

$$\Phi(t_n, y_n; \delta) := \frac{p_{m-1,\delta}(\delta A) - I}{\delta} y_n + \bar{p}_{m-1,\delta}(\delta A) g(t_n) \tag{51}$$

and so by (37) we have

$$\begin{aligned} d(t,\delta) &= \frac{\exp(-\delta A) - p_{m-1}(\delta A)}{\delta} y(t) \\ &\quad + \frac{1}{\delta}\int_0^\delta \exp(-(\delta-\tau)A) g(t+\tau)\,\mathrm{d}\tau - \bar{p}_{m-1,\delta}(\delta A) g(t) \\ &= d_1(t,\delta) + d_2(t,\delta) + d_3(t,\delta) \end{aligned}$$

where

$$\begin{aligned} d_1(t,\delta) &:= \frac{\exp(-\delta A) - p_{m-1}(\delta A)}{\delta} y(t) \\ d_2(t,\delta) &:= \frac{1}{\delta}\int_0^\delta \exp(-(\delta-\tau)A) g(t+\tau)\,\mathrm{d}\tau - \frac{1}{\delta}\int_0^\delta \exp(-(\delta-\tau)A) g(t)\,\mathrm{d}\tau \\ d_3(t,\delta) &:= \frac{1}{\delta}\int_0^\delta \exp(-(\delta-\tau)A) g(t)\,\mathrm{d}\tau - \bar{p}_{m-1,\delta}(\delta A) g(t) \end{aligned}$$

As in previous theorem, a bound for $d_1(t,\delta)$ is given by

$$\|d_1(t,\delta)\| \leqslant k_1 \delta^{m-1} \|y\|_{C[0,T]} \tag{52}$$

For $d_2(t,\delta)$, applying Lagrange's theorem to all components of g, we easily get

$$\|d_2(t,\delta)\| \leqslant k_3 \delta \|g'\|_{C[0,T]} \tag{53}$$

For $d_3(t,\delta)$, using (40) and (45) we find

$$\begin{aligned} \|d_3(t,\delta)\| &= \|(\varphi(\delta A) - \bar{p}_{m-1,\delta}(\delta A)) g(t)\| \\ &\leqslant k_4 \delta^{m-1} \|g\|_{C[0,T]} \end{aligned} \tag{54}$$

Finally, by (52)–(54), it follows that the method is consistent with the problem (1) with order 1. If in particular $g(t) = g$, clearly $d_2(t,\delta) = 0$ and so $\|d(t,\delta)\| \leqslant k_5 \delta^{m-1}$. □

Regarding the convergence of the above methods, we can state the following result.

Theorem 6.3
Let $m \geqslant 2$. If the F$[m,k]$ (or F$[m]$) method is consistent with order q, then it is convergent with the same order.

Proof
It is sufficient to show that the function $\Phi(t,y;\delta)$ defined by (47) (or (51)) is Lipschitzian with respect to the variable y. We have

$$\Phi(t,y_a;\delta) - \Phi(t,y_b;\delta) = \frac{p_{m-1,\delta}(\delta A) - I}{\delta}(y_a - y_b)$$

hence, using (45),

$$\begin{aligned}\|\Phi(t,y_a;\delta)-\Phi(t,y_b;\delta)\| &\leqslant \frac{1}{\delta}\,\|p_{m-1,\delta}(\delta A)-I\|\|y_a-y_b\|\\ &\leqslant \frac{1}{\delta}\,(\|p_{m-1,\delta}(\delta A)-\mathrm{e}^{-\delta A}\|+\|\mathrm{e}^{-\delta A}-I\|)\|y_a-y_b\|\\ &\leqslant \frac{1}{\delta}\left(c\delta^m+\sum_{j=1}^{\infty}\frac{\delta^j\|A\|^j}{j!}\right)\|y_a-y_b\|\\ &\leqslant (\|A\|+O(\delta))\|y_a-y_b\|\end{aligned}$$

Note that we did not distinguish between F$[m,k]$ and F$[m]$, because the proof is the same. □

7. SOME NUMERICAL CONSIDERATION

The practical implementation of the recursion (15) requires the evaluation of a certain number of the Laurent coefficients $\{\alpha_j\}_{j\geqslant 0}$ of the mapping ψ relative to a compact subset Ω containing $\sigma(A)$. In general, $\sigma(A)$ is not known and it is necessary to estimate it using some eigenvalue method (see e.g. References [7, 8, 20, 21]). Consequently, Ω is usually defined as the convex polygonal compact obtained joining the outermost points of the estimate set [20].

For the evaluation of the Laurent coefficients $\{\alpha_j\}_{j\geqslant 0}$ of ψ, in our tests we employed the software SC Matlab Toolbox, written by Driscoll in 1995 (see Reference [22]). We must point out that especially when Ω is convex, a small number of computed leading coefficients of ψ usually allows to get a good approximation of Ω. In Figure 1, we can see some examples of convex polygonal domains and the approximations obtained computing l (defined at the top of each picture) leading coefficients of the corresponding ψ.

In general, always assuming that Ω is convex, the choice of $l=10$ usually leads to good results. Hence, numerically, the mapping ψ with its (generally infinite) development is replaced by an approximation

$$\tilde{\psi}(w)=w+\tilde{\alpha}_0+\frac{\tilde{\alpha}_1}{w}+\frac{\tilde{\alpha}_2}{w^2}+\cdots+\frac{\tilde{\alpha}_l}{w^l},\quad |w|>\tilde{\gamma}$$

with $\tilde{\gamma}\approx\gamma$.

The implementation of the recursion (15) requires also the evaluation of the Faber coefficients $a_j(\delta)$, $j\geqslant 0$. After computing $\tilde{\psi}$, the computation of the Faber coefficients can be performed using any suitable quadrature formula. Anyway one can also use the following result.

Proposition 7.1
For $j\geqslant 1$,

$$a_{j-1}(\delta)+\frac{j}{\delta}\,a_j(\delta)=\textstyle\sum_{i=1}^{\infty}\mathrm{i}\alpha_i a_{j+i}(\delta) \tag{55}$$

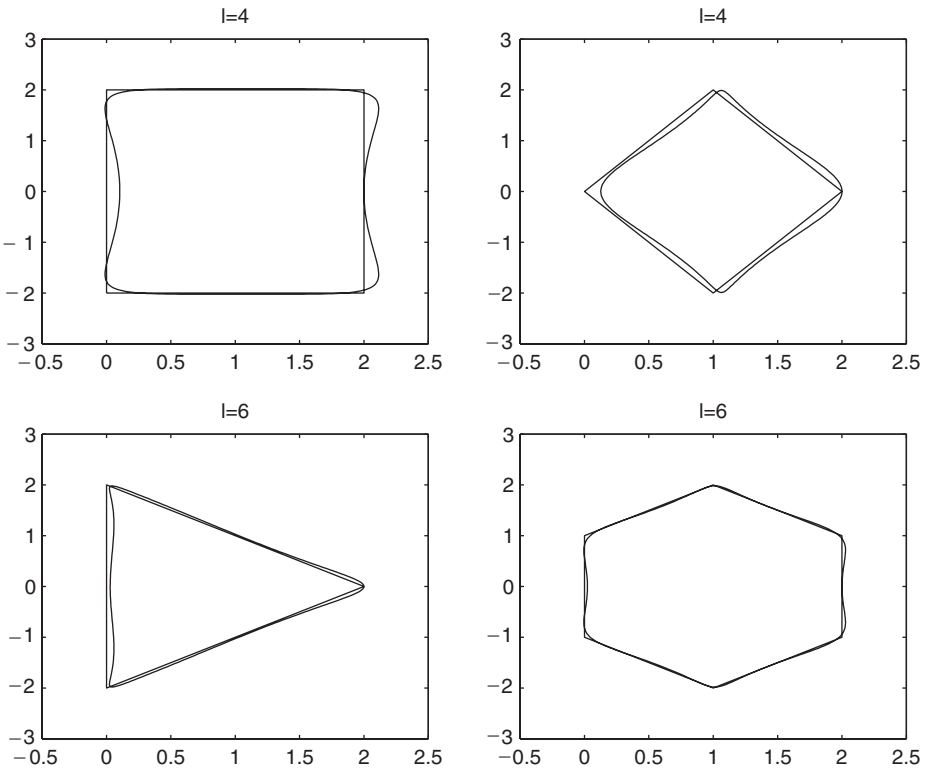

Figure 1.

Proof
Integrating by parts $a_j(\delta)$, $j \geqslant 1$, and by (10), we get

$$a_j(\delta) = \frac{1}{2\pi \mathrm{i}} \int_{|w|=R} \frac{\exp(-\delta\psi(w))}{w^{j+1}}\,\mathrm{d}w, \quad R > \gamma$$

$$= -\frac{\delta}{j}\,\frac{1}{2\pi \mathrm{i}} \int_{|w|=R} \frac{\exp(-\delta\psi(w))}{w^{j}}\psi'(w)\,\mathrm{d}w$$

$$= -\frac{\delta}{j}\,\frac{1}{2\pi \mathrm{i}} \int_{|w|=R} \frac{\exp(-\delta\psi(w))}{w^{j}}\left[1 - \frac{\alpha_1}{w^2} - \frac{2\alpha_2}{w^3} - \cdots\right]\mathrm{d}w$$

that proves the thesis. □

Clearly, on replacing ψ by $\tilde{\psi}$, formula (55) has a finite number $l+2$ of terms. It generalizes the well-known recursion for Chebyshev coefficients that can be expressed in terms of modified Bessel functions (see e.g. Reference [13]). As for Chebyshev coefficients, formula

(55) cannot be used forward, in view of its instability. Once *a priori* estimate of the required number m of coefficients is available (using for instance *a priori* upper bound for the error of approximation), after computing $a_{m+1}(\delta), a_{m+2}(\delta) \ldots, a_{m+j}(\delta)$, recursion (55) must be used backward. We observe that for both $\mathrm{F}[m,k]$ and $\mathrm{F}[m]$, the mapping $\tilde{\psi}$ and the Faber coefficients have to be computed only once, at the beginning.

Together with the above numerical consideration, here we want to show an improvement of the bound (22) for the exponential function, i.e.

$$|a_j(\delta)| \leqslant \frac{\mathrm{e}^{-\delta\psi(-R)}}{R^j} \tag{56}$$

Proposition 7.2
Let $\Omega \in \bar{\mathbf{M}}$. For each $R > \gamma$ we have

$$\begin{aligned} |a_0(\delta)| &\leqslant \mathrm{e}^{-\delta\psi(-R)} \\ |a_j(\delta)| &\leqslant \frac{\delta}{jR^{j-1}} \mathrm{e}^{-\delta\psi(-R)} \left(1 + \left(\frac{\gamma}{R}\right)^2\right), \quad j \geqslant 1 \end{aligned} \tag{57}$$

Proof
For $j = 0$ the thesis follows immediately by (13) and the geometry of Ω. For $j \geqslant 1$, using integration by parts we get

$$\begin{aligned} a_j(\delta) &= \frac{1}{2\pi\mathrm{i}} \int_{|u|=R} \frac{\exp(-\delta\psi(u))}{u^{j+1}} \, \mathrm{d}u \\ &= -\frac{\delta}{j} \frac{1}{2\pi\mathrm{i}} \int_{|u|=R} \frac{\exp(-\delta\psi(u))}{u^j} \psi'(u) \, \mathrm{d}u \end{aligned} \tag{58}$$

Now, using the bound (see Reference [11])

$$|\psi'(R\mathrm{e}^{\mathrm{i}\theta})| \leqslant 1 + \left(\frac{\gamma}{R}\right)^2, \quad R > \gamma$$

we get the thesis for $j \geqslant 1$. □

Defining $\Phi_j(\delta)$ and $\bar{\Phi}_j(\delta)$, $j \geqslant 0$, as the bounds for $|a_j(\delta)|$ given by (56) and (57), respectively, the following tables show a comparison between these two bounds for $\delta = 0.05$ and $\delta = 0.01$, in the case of an ellipse. In particular we consider the ellipse associated with the conformal mapping $\psi(w) = w + 5 + 2/w$, with capacity $\gamma = 4$. For the bounds we choose $R = \gamma$.

As we can see, the difference between the two estimates appears more evident as δ becomes smaller. The improvement given by (57) is of particular importance when $\delta \to 0$, because in this situation $\bar{\Phi}_0(\delta) \to 1$ and $\bar{\Phi}_j(\delta) \to 0$ for $j > 0$ (see (57)), as the corresponding exact values $|a_j(\delta)|$.

Note that Proposition 7.2 can be used to improve the bounds (21) and (26). However, such an improvement does not allow to attain qualitatively better results in Section 6.

j	$\lvert a_j(\delta)\rvert$	$\Phi_j(\delta)$	$\bar{\Phi}_j(\delta)$
0	7.8E-1	9.7E-1	9.7E-1
1	3.9E-2	2.4E-1	9.7E-2
2	6.5E-5	6.1E-2	1.2E-2
3	1.1E-9	1.5E-2	2.0E-3
4	$<$1E-14	3.8E-3	3.8E-4
5	$<$1E-14	9.5E-4	7.6E-5
0	9.5E-1	9.9E-1	9.9E-1
1	9.5E-3	2.5E-1	2.0E-2
2	6.3E-7	6.2E-2	2.5E-3
3	8E-14	1.5E-2	4.1E-4
4	$<$1E-14	3.9E-3	7.7E-5
5	$<$1E-14	9.7E-4	1.5E-5

8. NUMERICAL EXPERIMENTS

The problems we consider arise from the semi-discretization of the parabolic equation

$$\frac{\partial u(x,y,z,t)}{\partial t} = \Delta u(x,y,z,t) - \gamma_1 \frac{\partial u(x,y,z,t)}{\partial x} - \gamma_2 \frac{\partial u(x,y,z,t)}{\partial y} + r(x,y,z,t)$$

$$x, y, z \in (0,1), \quad \gamma_1, \gamma_2 \in \mathbf{R}$$

$$u(x,y,z,t) = 0 \quad \text{for } (x,y,z) \text{ on the boundary}$$

where Δ is the three-dimensional Laplacian operator, using central differences on a uniform meshgrid of $n+2$ points in each direction. The semi-discretization yields usual systems of ordinary linear differential equations of the type

$$\begin{aligned} y'(t) &= -Ay(t) + g(t), \quad t \in [0,T] \\ y(0) &= y_0 \end{aligned} \tag{59}$$

where $A \in \mathbf{R}^{N\times N}$, with $N = n^3$, and $g : [0,T] \to \mathbf{R}^N$. Defining $h = 1/(n+1)$, and

$$\lambda_n := \cos\left(\frac{\pi}{n+1}\right)\left(\sqrt{1-\frac{\gamma_1^2 h^2}{4}} + \sqrt{1-\frac{\gamma_2^2 h^2}{4}} + 1\right)$$

$\sigma(A)$ is contained in the rectangle $(1/h^2)R_n$, where

$$R_n := [6 - 2\,\mathrm{Re}\,\lambda_n, 6 + 2\,\mathrm{Re}\,\lambda_n] \times [-2\mathrm{i}\,\mathrm{Im}\,\lambda_n, 2\mathrm{i}\mathrm{Im}\,\lambda_n].$$

In order to provide an estimate for $F(A)$ of this example we use the following result [23, p. 79].

Theorem 8.1
Given $A \in \mathbf{R}^{N\times N}$, let $M := \frac{1}{2}(A + A^T)$, $\tilde{M} := \frac{1}{2}(A - A^T)$. Then

$$F(A) \subseteq [a,b] \times [-\mathrm{i}c, \mathrm{i}c] \tag{60}$$

where a, b, c, are such that $F(M)=[a,b]$, $F(\tilde{M})=[-ic,ic]$. The rectangle (60) is the smallest that contains $F(A)$.

In our case, M is the matrix obtained discretizing as before the Laplacian operator, so that

$$F(M)=\frac{6}{h^2}\left[1-\cos\frac{\pi}{n+1},1+\cos\frac{\pi}{n+1}\right]$$

and for $\tilde{M}$, we have

$$F(\tilde{M})=\frac{2(\gamma_1+\gamma_2)}{h}\left[-i\cos\frac{\pi}{n+1},i\cos\frac{\pi}{n+1}\right]$$

Therefore, once the conformal mapping ψ relative to $(1/h^2)R_n$ has been computed, the smallest value s such that $\Omega(s)\supseteq F(A)$, can be estimated by solving

$$\psi(w)=\frac{6}{h^2}\left(1-\cos\frac{\pi}{n+1}\right)+i\,\frac{2(\gamma_1+\gamma_2)}{h}\cos\frac{\pi}{n+1} \tag{61}$$

If $\tilde{w}$ is the solution of (61), then $s\leqslant|\tilde{w}|$.

In the following numerical examples we make a comparison between the Faber exponential integrators (39) and (41) built on the compact $\Omega:=(1/h^2)R_n$ and some Krylov exponential integrators. In particular we consider the standard Arnoldi exponential integrator, extensively studied in References [1, 6], implemented with the modified Gram–Schmidt process without reorthogonalization. We also consider the exponential integrator based on the incomplete Arnoldi process of order P, that we denote with Arnoldi (P). For this method, various numerical experiments on our example reveal that the choice of $P=4$ is the most convenient in the sense of the convergence rate with respect to the total amount of work. Finally, we consider the exponential integrator based on the Lanczos biorthogonalization algorithm, studied in References [2, 3]. Since all these approaches are of polynomial type like the FSM, we can make an effective comparison simply by choosing, for each problem, corresponding integrators (cf. Section 5 and Reference [1]). We integrate equations of type (59) between 0 and T, varying the time step δ. The degree m of the two methods at each step is chosen so that the final error, i.e. the error at time T (err in the tables below), is almost equal for all the methods. Our aim is to give particular attention to the computational costs of the methods relative to similar accuracy results. Computational costs are considered in terms of number of scalar products (nsp). In this context, we must point out that an application of A costs about as much as seven scalar products. In order to understand how much work would be required for problems with a less sparse matrix, we remark that the number of matrix–vector products is equal to the degree m for the Faber and the Arnoldi based exponential integrators, whereas it is equal to $2m$ for the Lanczos-based method.

In all tests Faber methods are built computing only the first four leading Laurent coefficients of the mapping ψ relative to $(1/h^2)R_n$, i.e., we approximate ψ by

$$\tilde{\psi}(w)=w+\tilde{\alpha}_0+\frac{\tilde{\alpha}_1}{w}+\frac{\tilde{\alpha}_2}{w^2}+\frac{\tilde{\alpha}_3}{w^3}$$

so that the corresponding recurrence relation for Faber polynomials involves six terms (cf. Section 3).

Table I.

				Arnoldi			Faber		
γ_1	γ_2	T	δ	m	nsp	err	m	nsp	err
50	20	0.05	0.05	50	**1605**	2.20E-9	70	**455**	1.57E-9
			0.025	40	**2168**	2.99E-9	36	**462**	3.34E-9
			0.01	18	**1449**	2.49E-8	18	**561**	1.85E-9
			0.005	12	**1572**	4.21E-9	11	**660**	2.84E-9
70	50	0.02	0.02	56	**1965**	8.51E-9	62	**402**	6.61E-9
			0.01	38	**1983**	4.62E-9	38	**488**	6.81E-9
			0.005	26	**2090**	5.78E-9	25	**633**	6.42E-9
			0.002	15	**2190**	2.69E-9	17	**1056**	6.28E-9
100	100	0.02	0.02	80	**3768**	1.30E-9	85	**554**	1.31E-9
			0.01	48	**2985**	1.44E-9	47	**607**	1.05E-9
			0.005	28	**2363**	5.91E-9	26	**660**	4.93E-9
			0.002	16	**2416**	1.14E-9	15	**924**	1.16E-9
				Arnoldi (4)			Lanczos		
50	20	0.05	0.05	75	**789**	2.92E-9	65	**988**	2.36E-9
			0.025	38	**793**	4.18E-9	40	**1216**	0.24E-9
			0.01	20	**1030**	1.44E-9	25	**1900**	6.20E-9
			0.005	12	**1212**	2.59E-9	14	**2128**	1.67E-8
70	50	0.02	0.02	63	**661**	6.49E-9	60	**912**	8.49E-9
			0.01	38	**793**	6.94E-9	37	**1124**	2.98E-9
			0.005	24	**993**	3.86E-9	30	**1824**	2.38E-9
			0.002	14	**1424**	2.04E-9	16	**2432**	4.96E-9
100	100	0.02	0.02	83	**873**	5.41E-9			
			0.01	48	**1005**	2.92E-9			
			0.005	29	**1205**	4.22E-9	32	**1945**	1.75E-8
			0.002	16	**1663**	1.44E-9	18	**2736**	3.33E-9

Remark 8.1

For our test problem, we use a matrix whose spectrum is explicitly known. Anyway, it is worth noting that we do not use the spectral decomposition of the matrix, but only the convex hull of its spectrum.

8.1. A non-symmetric model problem

In the first test problem we define $n=15$, i.e. $N=3375$, $h=1/16$, $g\equiv 0$, $y_0=(1,1,\ldots,1)^{\mathrm{T}}$ and vary γ_1, γ_2. Since $g\equiv 0$, the problem simply consists of computing a certain number of matrix exponentials. Hence we compare the F[m] method with the corresponding Arnoldi, Arnoldi (4) and Lanczos schemes.

In Table I we can note that even if, in most of the tests, the number of iterations m of the Krylov methods is less than the corresponding m of the Faber method, in terms of the number of scalar products the F[m] method performs surely better. This difference of costs is particularly evident with respect to the Arnoldi method for large (with respect to T) values of δ, because in this situation, to get a certain accuracy, m has to be chosen sufficiently large. In fact, as well known, the cost of each Arnoldi step increases with the number of iterations.

Table II.

					Arnoldi			Faber	
γ_1	γ_2	T	δ	m	nsp	err	m	nsp	err
60	0	0.1	0.1	50	**3210**	1.11E-9	66	**858**	1.64E-9
			0.05	48	**4478**	3.07E-9	64	**1247**	3.48E-9
			0.02	40	**6504**	6.31E-9	46	**1782**	4.74E-9
			0.01	30	**7293**	1.70E-9	34	**2395**	2.58E-9
			0.005	21	**7761**	2.11E-9	26	**3465**	1.12E-9
40	20	0.1	0.1	47	**2876**	4.84E-9	55	**713**	5.25E-9
			0.05	46	**4153**	9.50E-9	53	**1029**	6.01E-9
			0.025	44	**6402**	1.08E-9	50	**1617**	1.49E-9
			0.01	30	**7293**	1.60E-9	35	**2468**	0.64E-9
			0.005	20	**7182**	3.19E-9	25	**3326**	4.03E-9
					Arnoldi (4)			Lanczos	
60	0	0.1	0.1	70	**1472**	1.11E-9	60	**1824**	2.57E-9
			0.05	68	**2144**	3.07E-9			
			0.02	60	**3780**	6.31E-9			
			0.01	31	**3548**	1.70E-9	30	**5016**	2.43E-8
			0.005	22	**4771**	2.11E-9	21	**6073**	1.63E-8
40	20	0.1	0.1	57	**1196**	4.84E-9	55	**1672**	8.63E-9
			0.05	57	**1794**	9.50E-9			
			0.025	48	**2514**	1.08E-9			
			0.01	31	**3548**	1.60E-9			
			0.005	21	**4548**	3.19E-9	23	**7341**	5.75E-9

For this reason, especially when m is large, the incomplete version Arnoldi(4) represents an effective schemes. Regarding the Lanczos method, the experiments show that sometimes it does not allow to attain the accuracy of the other methods. Moreover, the empty lines in the table below indicates that the Lanczos method is very unstable and does not converge (the same is valid for the Tables II and III).

8.2. A non-symmetric model problem with forcing term

As before we define $n = 15$, i.e. $N = 3375$, $y_0 = (1, 1, \ldots, 1)^{\mathrm{T}}$. In Table II some tests on a problem (59) with time constant forcing term $g = y_0$, are shown. Regarding the computational cost, all consideration given for the previous example remains valid also for this one. In this example we use iteration (43) for all the methods and, in order to optimize the cost, it is also possible to distinguish between the number of iterations performed to compute the matrix exponential (m_1) and that for the computation of $\varphi(\delta A)$ (m_2). In any case, the relationship between the costs of the methods is substantially independent of this choice.

In Table III we can observe some tests on a problem (59) with the time varying forcing term $g(t) = y_0 t$. Since we use the scheme described by (39) we need a quadrature formula. The formula chosen (q in Table III) is the Simpson rule built on k points that we indicate with S_k. Therefore, the weights $(\mu_1, \mu_2, \mu_3, \ldots, \mu_k)$ of (39) are given by $(\frac{1}{3}, \frac{4}{3}, \frac{2}{3}, \ldots, \frac{1}{3})$.

Table III.

					Arnoldi			Faber		
γ_1	γ_2	T	δ	q	m	nsp	err	m	nsp	err
40	0	0.1	0.1	S_9	20	**2736**	7.29E-3	25	**1267**	7.32E-3
			0.05	S_7	25	**6370**	3.28E-3	30	**2488**	3.42E-3
			0.025	S_7	16	**6523**	2.14E-3	25	**4276**	2.66E-3
			0.02	S_7	16	**8214**	1.25E-3	25	**5385**	1.61E-3
			0.01	S_5	20	**16758**	1.26E-4	25	**7761**	1.27E-4
					Arnoldi (4)			Lanczos		
40	0	0.1	0.1	S_9	26	**2156**	5.97E-3	20	**2432**	3.81E-3
			0.05	S_7	30	**4056**	5.61E-3	25	**4940**	5.64E-3
			0.025	S_7	18	**4989**	7.14E-3	16	**6566**	7.22E-2
			0.02	S_7	17	**5922**	7.07E-3	16	**8268**	4.13E-2
			0.01	S_5	20	**10094**	7.24E-3	18	**13406**	7.13E-3

As we can see, the behaviour is similar to that of previous examples. It is not much useful to give more complicated tests, because all the methods are applied in the same manner and the only important difference is given by the computation of the matrix functions.

9. FINAL REMARKS

In our tests we are supposed to know the exact location of the spectrum of the matrix A, that is, the rectangle $(1/h^2)R_n$. However, as explained in Section 7, when $\sigma(A)$ is not known the implementation of the Faber expansion method requires a preliminary work to locate it. Some authors consider such preliminary phase as a serious drawback of expansion methods for matrix functions. Anyway, for practical problems like (1), where more than one or a lot of matrix functions always with the same matrix must be computed, the cost of the preliminary phase is surely negligible with respect to the total amount of work (see e.g. References [8, 13]). Moreover, other numerical experiments show that when the function involved is analytic in the whole complex plane as in (1), even a poor approximation of $\sigma(A)$ leads to acceptable results.

For these reasons, in the author's opinion, the Faber expansion method actually constitutes an efficient alternative to the Krylov approaches for (1) based on the Arnoldi or Lanczos algorithms and their variants, especially when the matrix of the problem is large and sparse as in the case considered for the numerical experiments.

REFERENCES

1. Gallopoulos E, Saad Y. Efficient solution of parabolic equations by Krylov approximation methods. *SIAM Journal on Statistical and Scientific Computing* 1992; **13**:1236–1264.
2. Hochbruck M, Lubich C. On Krylov subspace approximation to the matrix exponential operator. *SIAM Journal on Numerical Analysis* 1995; **34**(5):1911–1925.
3. Hochbruck M, Lubich C. Exponential integrators for quantum-classical molecular dynamics. *BIT* 1999; **39**: 620–645.

4. Hochbruck M, Lubich C, Selhofer H. Exponential integrators for large systems of differential equations. *SIAM Journal on Scientific Computing* 1998; **19**:1552–1574.
5. Knizhnerman L. Calculation of functions of unsymmetric matrices using Arnoldi's method. *Computational Mathematics and Mathematical Physics* 1991; **31**(1):1–9 (English Edition by Pergamon Press: Oxford).
6. Saad Y. Analysis of some Krylov subspace approximations to the matrix exponential operator. *SIAM Journal on Numerical Analysis* 1992; **29**:209–228.
7. Moret I, Novati P. The computation of functions of matrices by truncated Faber series. *Numerical Functional Analysis and Optimization* 2001; **22**:697–719.
8. Novati P. Polynomial methods for the computation of functions of large unsymmetric matrices. *Ph.D. Thesis*, Università degli Studi di Padova, 2000.
9. Curtiss JH. Faber polynomials and Faber series. *American Mathematical Monthly* 1971; **78**:577–596.
10. Ellacott SW. Computation of Faber series with application to numerical polynomial approximation in the complex plane. *Mathematics of Computation* 1983; **40**:575–587.
11. Kovari T, Pommerenke C. On Faber polynomials and Faber expansions. *Mathematische Zeitschrift* 1967; **99**:193–206.
12. Smirnov VI, Lebedev NA. *Functions of a complex variable—constructive theory.* MIT Press: Cambridge, MA, 1968.
13. Bergamaschi L, Vianello M. Efficient computation of the exponential operator for large, sparse, symmetric matrices. *Numerical Linear Algebra with Applications* 2000; **7**(1):27–45.
14. Tal-Ezer H. Spectral methods in time for hyperbolic equations. *SIAM Journal on Numerical Analysis* 1986; **23**:11–26.
15. Tal-Ezer H. Spectral methods in time for parabolic problems. *SIAM Journal on Numerical Analysis* 1989; **26**:1–11.
16. Eiermann M. On semiiterative methods generated by Faber polynomials. *Numerische Mathematik* 1989; **56**:139–156.
17. Eiermann M, Niethammer W, Varga RS. A study of semiiterative methods for nonsymmetric systems of linear equations. *Numerische Mathematik* 1985; **47**:505–533.
18. Eiermann M. Fields of values and iterative methods. *Linear Algebra and Its Applications* 1993; **180**:167–197.
19. Spijker MN. Numerical ranges and stability estimates. *Applied Numerical Mathematics* 1993; **13**:241–249.
20. Starke G, Varga RS. A hybrid Arnoldi–Faber iterative method for nonsymmetric systems of linear equations. *Numerische Mathematik* 1993; **64**:213–240.
21. Manteuffel TA, Starke G. On hybrid iterative methods for nonsymmetric systems of linear equations. *Numerische Mathematik* 1996; **73**:489–506.
22. Driscoll TA. Algorithm 756: A MATLAB toolbox for Schwarz–Christoffel mapping. *ACM Transactions on Mathematical Software* 1996; **22**(2):168–186.
23. Householder AS. *The Theory of Matrices in Numerical Analysis.* Blaisdell: New York, 1964.

