A POLYNOMIAL METHOD BASED ON FEJÈR POINTS FOR THE COMPUTATION OF FUNCTIONS OF UNSYMMETRIC MATRICES

PAOLO NOVATI*

Abstract. In this paper we consider the problem of the computation of functions of large unsymmetric matrices by vectors. We study a polynomial method based on interpolation at the Fejèr points. We provide a detailed error analysis and apply the method to some examples that occur in practical problems. Finally, numerical experiments are shown.

1. Introduction. Let A be a $N \times N$ real nonsymmetric matrix and v a N-dimensional real vector. We consider the problem of the computation of

$$y = f(A)v, \tag{1.1}$$

where f is a given function that we suppose to be analytic in a certain domain of the complex plane containing the spectrum of A, that we denote with $\sigma(A)$. From the Cauchy Integral Theorem, $f(A) \in \mathbf{R}^{N\times N}$ is given by

$$f(A) = \frac{1}{2\pi i}\int_{\Lambda} f(z)(zI - A)^{-1}\, dz, \tag{1.2}$$

where Λ is the boundary curve of a piecewise smooth bounded region containing $\sigma(A)$ and where f is analytic.

An example of (1.1) is the solution of linear systems of equations, where $f(z) = 1/z$ or $f(z) = 1/(1-z)$. Moreover, (1.1) can also appear in some differential problems, such as solving semidiscrete approximations of partial differential equations of elliptic type, where $f(z) = \exp(-z^{1/2})$, of hyperbolic type, with $f(z) = \cos(z^{1/2})$, or of parabolic type, with $f(z) = \exp(z)$. The case of $f(z) = (\exp(z) - 1)/z$ is also important in the context of systems of linear initial value problems. The case of linear systems is supported by a seemingly unbounded literature, but also the other cases have been extensively studied (see for instance [3], [4], [6], [11], [13], [15], [19], [20], [21], [22]).

In particular, in recent years several people have proposed methods based on the projection of the operator $f(A)$ onto the Krylov subspaces generated by A and v, using the Arnoldi or Lanczos algorithms (see [3], [4], [11], [13], [22]). This kind of technique (for which we shall use the usual abbreviation KPMs, *Krylov Projection Methods*) is of polynomial type, that is, the approximations are of the type

$$y_m = p_{m-1}(A)\, v \approx f(A)\, v, \quad m \geq 1, \tag{1.3}$$

where $p_{m-1} \in \Pi_{m-1}$, the set of the polynomials of degree at most $m-1$. Concerning the rate of convergence, KPMs generally show a very satisfactory behavior (see [13]), but they also present the typical disadvantages of the projective schemes, that is, the construction of the projection subspaces. In detail, using the classical Arnoldi algorithm to build the Krylov subspaces, there is the well known problem of growth of computational cost. On the other hand, using the Lanczos algorithm, there is the possibility of total breakdown with consequent failure of the method, and in general there are stability problems due to the fact that oblique projections instead

*Università degli Studi di Trieste, Dipartimento di Scienze Matematiche, Via Valerio 12/1, 34100 TS.

of orthogonal projections are used (see [11], [22]). Another disadvantage of these methods is the strict dependence on v, in the sense that the Krylov subspaces are built with respect to this vector, which is a drawback when computing $f(A)v$ with different v, as for instance in the context of differential equations problems [11].

To overcome the drawbacks of KPMs, in this paper we consider a particular method (1.3) in which the polynomials p_{m-1} interpolate f at *Fejér points*, that are built using some a priori information about $\sigma(A)$. We call *Fejèr Points Method* (FPM) such a method. Assuming to know a compact subset $\Omega \subset \mathbf{C}$ such that $\sigma(A) \subseteq \Omega$, the idea is to define the interpolation points such that the corresponding interpolating sequence $\{p_{m-1}\}_{m\geq 1}$ satisfies

$$\|p_{m-1} - f\|_{\Omega} \to 0, \tag{1.4}$$

where $\|\cdot\|_{\Omega}$ is the supremum norm on Ω. As we shall see, the condition (1.4) is not always sufficient for the convergence of the corresponding polynomial method, i.e., (1.4) does not ensure

$$(p_{m-1}(A) - f(A))\, v \to 0.$$

A stronger sufficient condition is that the sequence $\{p_{m-1}\}_{m\geq 1}$ *converges maximally* to f on Ω and the choice of Fejér points ensures the realization of this last condition.

The most important features of the FPM are the following:

1. varying m the approximations y_m are obtained by means of a three term recurrence relation;
2. the parameters involved are independent of v;
3. the computational cost at each step is essentially that of an application of A.

On the other hand, the most important drawback of the FPM is that we need some a priori informations about $\sigma(A)$, that is, contrary to the KPMs, this method is not "matrix free". In fact, since it is based on the complex approximation of f in a compact Ω containing $\sigma(A)$, obviously we need to know the location of at least the outermost eigenvalues of A. Hence, to apply the FPM, we need a previous phase where an eigenvalue method is used. In literature, such composite procedures are usually called *hybrid methods*. However, we observe that in many practical problems, one needs $f(A)v$ for several vectors v, so that in these cases the initial cost for the approximation of $\sigma(A)$ becomes negligible.

Other techniques based on the approximation of f on a certain Ω such that $\sigma(A) \subseteq \Omega$, have been already studied in the case of linear systems (*polynomial acceleration methods*, *semi-iterative methods*, see e.g. [7], [9], [17]), and in the case of the exponential function (see e.g. [1], [19], [20], [21], [27], [28]), where, depending on the geometry of Ω, the polynomials p_{m-1} are defined as truncated Chebyshev or Faber series with respect to f and Ω in order to obtain maximal convergence.

The use of Fejèr points for solving problems of type (1.1) had been already studied in [9] in order to get a *first-order Richardson method* for solving linear systems. However, in [9] neither error analysis nor numerical experiments were given. Herein, we extend the arguments already studied for the function $1/z$ (involved in the context of linear systems) to the general case of f and provide a detailed error analysis. In this analysis, we shall also distinguish between the cases of f analytic in the whole complex plane and f singular at some finite point. Error bounds for some f of practical relevance are given. At the end, we shall also show the effectiveness of this method

on a number of experiments.

The paper is organized as follows: in §2 we give a general convergence analysis of polynomial methods, with particular attention to the concept of asymptotically optimal methods. In §3, we study the possibility of using interpolatory approaches for (1.1) and consider the particular choice of Fejér points as interpolation nodes and their properties. In §4 we provide an error analysis of the FPM and in §5 we consider some examples occurring in practical problems. §6 is devoted to the explanation of some important computational details; in this section we also provide the algorithm that is at the basis of the hybrid procedures that we test in §7.

2. Convergence analysis of polynomial methods. First, let us introduce the following class of compact subsets of the complex plane,

$$\mathbf{M} := \left\{ \Omega \subset \mathbf{C} : \Omega \text{ is compact, } \overline{\mathbf{C}} \backslash \Omega \text{ is simply connected and } \Omega \text{ contains infinitely many points} \right\}.$$

Given $\Omega \in \mathbf{M}$, by the Riemann Mapping Theorem there exists a conformal surjection

$$\psi : \overline{\mathbf{C}} \setminus \{w : |w| \leq \gamma\} \to \overline{\mathbf{C}} \setminus \Omega,$$

such that $\psi(\infty) = \infty$, $\psi'(\infty) = 1$. The constant γ is called the *capacity* of Ω. By conformal mapping theory, it is known that the conformal surjection ψ has a Laurent expansion of the type

$$\psi(w) = w + \alpha_0 + \frac{\alpha_1}{w} + \frac{\alpha_2}{w^2} + ..., \qquad |w| > \gamma. \tag{2.1}$$

Let $\phi : \overline{\mathbf{C}} \setminus \Omega \to \overline{\mathbf{C}} \setminus \{w : |w| \leq \gamma\}$ be the inverse mapping of ψ. Moreover, let Γ be the boundary of Ω and, for $r > \gamma$, let $\Gamma(r)$ be the equipotential curve

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

Let us denote by $\Omega(r)$ the bounded domain with boundary $\Gamma(r)$. Clearly, $\Omega(\gamma) = \Omega$ and $\Gamma(\gamma) = \Gamma$.

Let f be a complex valued function analytic on $\Omega \in \mathbf{M}$. For the construction of a polynomial method for (1.1), we are mainly interested in searching sequences of polynomials $\{p_{m-1}\}_{m \geq 1}$ such that

$$\lim_{m \to \infty} \|f - p_{m-1}\|_\Omega = 0. \tag{2.2}$$

Now, consider a general method (1.3). Define the error of the m-th approximation $y_m = p_{m-1}(A)\, v$ as

$$e_m := f(A)v - p_{m-1}(A)\, v, \qquad m \geq 0. \tag{2.3}$$

If A is diagonalizable by X, and if $\sigma(A) \subseteq \Omega$, then for the Euclidean norm of the error (2.3) we have

$$\begin{aligned} \|e_m\|_2 &= \|f(A)v - p_{m-1}(A)\, v\|_2 \\ &\leq \|X\|_2 \left\|X^{-1}\right\|_2 \max_{\lambda \in \sigma(A)} |f(\lambda) - p_{m-1}(\lambda)| \, \|v\|_2 \\ &\leq \|X\|_2 \left\|X^{-1}\right\|_2 \|f - p_{m-1}\|_\Omega \, \|v\|_2 , \end{aligned}$$

so that, by condition (2.2), the convergence is ensured. If A is not diagonalizable, the condition (2.2) is generally not sufficient for the convergence of $\{y_m\}_{m\geq 1}$, so that we need a stronger uniform convergence condition.

If $\left\{p^*_{m-1}\right\}_{m\geq 1}$ is the sequence of polynomials of best uniform approximation of f on Ω, we want the sequence $\{p_{m-1}\}_{m\geq 1}$ to satisfy the condition

$$\overline{\lim_{m\to\infty}} \|p_{m-1}-f\|_\Omega^{1/(m-1)} = \overline{\lim_{m\to\infty}} \left\|p^*_{m-1}-f\right\|_\Omega^{1/(m-1)} =: k(\Omega,f), \tag{2.4}$$

that is, we want that $\{p_{m-1}\}_{m\geq 1}$ behaves asymptotically as the best polynomial approximation sequence. The quantity $k(\Omega,f)$ is usually called *asymptotic convergence factor*. Property (2.4) is known as *maximal convergence* of $\{p_{m-1}\}_{m\geq 1}$, and this sequence is usually called *asymptotically optimal* with respect to f and Ω.

If $\overset{\circ}{\Omega}(r)$ denotes the interior of $\Omega(r)$, we have that the quantity

$$\eta = \eta(f) := \max\left\{r : r>\gamma,\ f \text{ analytic on } \overset{\circ}{\Omega}(r)\right\}, \tag{2.5}$$

gives a measure of the optimal rate of convergence, because we have (cf. [31] Ch.4, Th.76)

$$k(\Omega,f) = \frac{\gamma}{\eta}. \tag{2.6}$$

By (2.6), for functions analytic in the whole complex plane, the rate of convergence of asymptotically optimal sequences is superlinear, because η can be chosen arbitrarily large.

As consequence of the maximal convergence property, we have the following result, that will be fundamental in the sequel, and that defines the so called *overconvergence*.

THEOREM 2.1. *([31], Ch.4, Th.6) Given $\Omega \in \mathbf{M}$, if $\{p_{m-1}\}_{m\geq 1}$ converges maximally to f on Ω, then this sequence converges uniformly on compact subsets of $\overset{\circ}{\Omega}(\eta)$.*

The following theorem recalls some criteria for convergence of a polynomial method. ∎

THEOREM 2.2. *[12] Let $m(z)=\prod_{i=1}^{\nu}(z-\lambda_i)^{n_i}$ be the minimal polynomial of A. The sequence $y_m = p_{m-1}(A)\,v$ converges to $f(A)v$ for each v if and only if*

$$\lim_{m\to\infty} p^{(j)}_{m-1}(\lambda_i) = \left[\frac{d^j}{dz^j} f(z)\right]_{z=\lambda_i}, \quad 1\leq i\leq \nu, \quad 0\leq j\leq n_i-1. \tag{2.7}$$

In case of diagonalizable A the criterion (2.7) reduces to

$$\lim_{m\to\infty} p_{m-1}(\lambda) = f(\lambda), \qquad \text{for each } \lambda\in\sigma(A).$$

Moreover, if $G\subset\mathbf{C}$ is an open set such that $\sigma(A)\subset G$ and such that f is analytic in G, then the following condition is sufficient

$$\{p_{m-1}\}_{m\geq 1} \quad \begin{array}{l} \text{converges to } f, \text{ uniformly} \\ \text{on each compact subset of } G. \end{array} \tag{2.8}$$

Hence, by Theorem 2.1 and condition (2.8) the following result holds.

THEOREM 2.3. *If* $\{p_{m-1}\}_{m\geq 1}$ *converges maximally to* f *on* Ω*, and if* $\sigma(A) \subset \overset{\circ}{\Omega}(\eta)$*, then* $p_{m-1}(A)\,v$ *converges to* $f(A)v$*.*

For the *asymptotic convergence factor* of the method, defined as

$$\overline{\lim_{m\to\infty}} \left[\|e_m\|^{1/m}\right],$$

since maximal convergence persists after differentiation (cf.[31], Ch.4, Th.7), we have the following result, which is a direct consequence of (2.6).

THEOREM 2.4. *[9] If* $\{p_{m-1}\}_{m\geq 1}$ *converges maximally to* f *on* Ω *and* $\sigma(A) \subseteq \Omega(r)$ *for* $\gamma \leq r < \eta$*, then*

$$\overline{\lim_{m\to\infty}} \left[\|e_m\|^{1/m}\right] \leq \frac{r}{\eta}. \tag{2.9}$$

By (2.9), for functions analytic in the whole complex plane, the rate of convergence of the method is superlinear, because η can be taken arbitrarily large.

A polynomial method based on a sequence $\{p_{m-1}\}_{m\geq 1}$ of Theorem 2.4 is said to be *asymptotically optimal* with respect to f and Ω.

3. The method based on Fejér points. Let $f : \mathbf{C} \to \mathbf{C}$. Given an integer m and a set $\{\xi_i^{(m)}\}_{i=1,\ldots,m}$ of distinct points in the complex plane, let $p_{m-1} \in \Pi_{m-1}$ be the polynomial that interpolates f at these points:

$$\begin{gathered} p_{m-1}(z) = \textstyle\sum_{k=1}^{m} f\left(\xi_k^{(m)}\right) \frac{\omega_m(z)}{\left(z-\xi_k^{(m)}\right)\omega_m'\left(\xi_k^{(m)}\right)}, \\ \omega_m(z) := \left(z-\xi_1^{(m)}\right)\cdot\ldots\cdot\left(z-\xi_m^{(m)}\right). \end{gathered} \tag{3.1}$$

The sequence $\{p_{m-1}\}_{m\geq 1}$ can be used to build a polynomial interpolation method for $f(A)v$. When $\xi_i = \xi_i^{(m)}$, the polynomials (3.1) can be written in Newton form

$$p_{m-1}(z) = f_0 + f_1(z-\xi_1) + \ldots + f_{m-1}(z-\xi_1)\cdot\ldots\cdot\left(z-\xi_{m-1}\right), \tag{3.2}$$

where f_k $(k = 0, \ldots, m-1)$ are the usual Newton coefficients:

$$f_0 = f[\xi_1], \qquad f_k := f\left[\xi_1, \ldots, \xi_{k+1}\right], \quad k = 1, \ldots, m-1. \tag{3.3}$$

Thus, if $\xi_i = \xi_i^{(m)}$, the sequence (3.2) of interpolatory polynomials fullfils the recurrence

$$p_{m-1}(z) = p_{m-2}(z) + \frac{f_{m-1}}{f_{m-2}}\left(z-\xi_{m-1}\right)\left(p_{m-2}(z) - p_{m-3}(z)\right), \quad m \geq 1. \tag{3.4}$$

It is easy to show that for the corresponding polynomial interpolation method the following 3-terms recurrence relation holds

$$\begin{gathered} y_0 = 0, \qquad y_1 = f_0 v, \\ y_m = y_{m-1} + \tfrac{f_{m-1}}{f_{m-2}}\left(A - \xi_{m-1} I\right)\left(y_{m-1} - y_{m-2}\right), \quad m \geq 2. \end{gathered} \tag{3.5}$$

The issue is to choose the interpolation nodes such that the corresponding method is convergent. The following theorem gives necessary and sufficient conditions for having polynomial sequences which converge maximally to f on a certain $\Omega \subset \mathbf{C}$.

THEOREM 3.1. *(Basic Convergence Principle, [24]) Let $\Omega \subset \mathbf{M}$. Suppose that the sequence of nodes $\{\xi_i\}_{i\geq 1}$ is entirely contained in Ω or that all its limit points are not external to Ω. Then for each function f, analytic in Ω, the sequence of interpolatory polynomials $\{p_{m-1}\}_{m\geq 1}$ is such that*

$$\lim_{m\to\infty} \|p_{m-1} - f\|_\Omega = 0,$$

if and only if

$$\lim_{m\to\infty} |(z-\xi_1)\cdot ... \cdot (z-\xi_m)|^{1/m} = \gamma(\Omega)\,|\phi(z)|\,, \tag{3.6}$$

uniformly in each compact subset of $\overline{\mathbf{C}}\backslash\Omega$. If this condition is fulfilled then $\{p_{m-1}\}_{m\geq 1}$ converges maximally to f on Ω.

In order to introduce the Fejèr points, let us define

$$\widetilde{\mathbf{M}} := \{\Omega \in \mathbf{M} : \phi \text{ can be extended to } \Gamma\}\,.$$

Given $\Omega \in \widetilde{\mathbf{M}}$, if u_j is the j-th root of unity, then the points

$$\xi_j := \psi(\gamma u_j)\,, \qquad j \geq 1, \tag{3.7}$$

are defined *Fejér points*. As proved in ([24] p.28, Th.1), the points (3.7) fullfils the relation (3.6). Hence, by Theorem 2.3, the corresponding method for computing $f(A)\,v$ is asymptotically optimal and converges if $\sigma(A) \subset \overset{\circ}{\Omega}(\eta)$, with η defined by (2.5). We call this method *Fejér Point Method* (FPM). The sequence of approximations can be iteratively obtained using (3.5).

REMARK 3.2. *Given $n = 2^k$ with k positive integer, it is easy to prove that, if Ω is a real bounded interval, than the points $\{\xi_j\}_{j=1...n}$ defined by (3.7) coincides with the Chebyshev-Gauss-Lobatto points.*

4. Error bounds. In this section we want to give bounds for the error e_m of the FPM. Since we want to deal with the general case of A not diagonalizable, we work with the *field of values* of A, defined as

$$F(A) := \left\{ \frac{z^H A z}{z^H z} : z \in \mathbf{C}/\{0\} \right\}. \tag{4.1}$$

For the following result see ([25] Th.4.1).

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

$$||(zI - A)^{-1}||_2 \leq 1/d(z, F(A)).$$

Using the above lemma we can prove the following result, that holds for each polynomial method.

THEOREM 4.2. *Let $\Omega \in \mathbf{M}$ and f analytic on Ω. Assume that $F(A) \subseteq \Omega(s)$, for some $\gamma \leq s < \eta$. Then, for any $s < r < \eta$,*

$$\|e_m\|_2 \leq \|v\|_2\, \|f - p_{m-1}\|_{\Omega(r)}\, \frac{(r+s)}{(r-s)}. \tag{4.2}$$

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

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

Then we get

$$\|e_m\|_2 \leq \frac{\|f - p_{m-1}\|_{\Omega(r)}}{2\pi} \int\limits_{|w|=r} \left|\psi'(w)\right| ||(\psi(w)I - A)^{-1}v||_2 dw.$$

Hence, by Lemma 4.1, we obtain

$$\|e_m\|_2 \leq \frac{\|v\|_2 \, \|f - p_{m-1}\|_{\Omega(r)}}{2\pi} \int\limits_{|w|=r} \left|\frac{\psi^{'}(w)}{\psi(w) - u}\right| dw,$$

where $u \in \Omega(s)$. Finally, since

$$\int\limits_{|w|=r} \left|\frac{\psi^{'}(w)}{\psi(w) - u}\right| dw = \frac{1}{r} \int\limits_{|w|=r} \left|\frac{w\psi^{'}(w)}{\psi(w) - u}\right| dw,$$

using the relation ([24] p.16)

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

the theorem is proved.
□

By Theorem 4.2, in order to get error bounds for the error of the FPM, it is necessary to bound the quantity $\|f - p_{m-1}\|_{\Omega(r)}$, for $r > \gamma$, where p_{m-1} interpolates f at Fejèr points. The bounds depend on the regularity of Γ. We have the following result, essentially a collection of results from [24].

THEOREM 4.3. *Let $\Omega \in \widetilde{\mathbf{M}}$ and f analytic on Ω. Then, for $\gamma \leq r < R < \eta$, the residual term of interpolation in Fejèr points has the value*

$$\|f - p_{m-1}\|_{\Omega(r)} \leq 2\,\|f\|_{\Gamma(R)}\, K_m \frac{R+r}{R-r} \frac{\left(\frac{r}{R}\right)^m}{1 - \left(\frac{r}{R}\right)^m}, \tag{4.3}$$

where $K_m \leq K$ for each m, is a quantity depending on the regularity of Γ:

1. if Γ has continuously turning tangent, $K = \frac{\rho_2}{\rho_1} e^{2\pi}$ and $K_m = K$ for each m, where ρ_1 and ρ_2 are such that

$$0 < \rho_1 \leq \frac{|\psi(t) - \psi(w)|}{|t - w|} \leq \rho_2;$$

2. if Γ is an analytical Jordan curve,

$$K_m = e^{\frac{A_n}{m^{2n-1}}}, \qquad n = 1, 2, ..., \tag{4.4}$$

where A_n depends only on n and Γ. If ψ is analytic in a domain $\gamma/s < |w| < +\infty$, $s > 1$, then we can put $A_n = A_n(s)$, where

$$A_n(s) := \frac{2}{3}\pi^2 \frac{\sqrt{(4n-1)!}\left(\frac{s}{\gamma}\right)^{2n-1}}{\left(\left(\frac{s}{\gamma}\right)^2 - 1\right)^{2n}} \sqrt{\ln \frac{\left(\frac{s}{\gamma}\right)^2}{\left(\frac{s}{\gamma}\right)^2 - 1}};$$

3. if ψ in (2.1) has a Laurent expansion with a finite number of terms, $K = 1$ and $K_m = K$ for each m.

Proof. For Γ with continuous turning tangent, the proof is given in [24] p.37 Lemma 4.

For Γ analytical Jordan curve, the proof is given in [24] p.38, Theorem 4.

For the last case, if ψ has a Laurent expansion with a finite number of terms, then ψ can be extended analytically to the domain $0 < |w| < +\infty$, and the constant $A_n(s)$ can be chosen for each $1 < s < +\infty$. Since $A_n(s)$ is a decreasing function of s and

$$\lim_{s\to\infty} A_n(s) = 0, \qquad n = 1, 2, ...,$$

the thesis follows by (4.4). □

The case of ψ with a Laurent expansion with a finite number of terms, is especially important from a computational point of view. In fact, even if there are infinitely many nonzero Laurent coefficients of ψ, it is obviously impossible to compute all of them. So, numerically, one always works with conformal mappings with a finite number of terms. This includes the important case of the three terms expansion of ψ, which is equivalent to consider compact subsets Ω whose boundary is an ellipse, or, in the degenerate case, Ω is an interval.

Using Theorems 4.2 and 4.3 we have the following result.

THEOREM 4.4. *Let $\Omega \in \mathbf{M}$. Assume that $F(A) \subseteq \Omega(s)$, for some $\gamma \le s < \eta$. Then for each r, R such that $s < r < R < \eta$, for the FPM we have*

$$\|e_m\|_2 \le 2\|f\|_{\Gamma(R)} K \frac{R+r}{R-r} \frac{\left(\frac{r}{R}\right)^m}{1-\left(\frac{r}{R}\right)^m} \frac{r+s}{r-s} \|v\|_2, \tag{4.5}$$

where K is defined in Theorem 4.3.

Next, we distinguish between f analytic in the whole complex plane and f singular at some finite point.

THEOREM 4.5. *Assume that $F(A) \subseteq \Omega(s)$. If f is analytic in the whole complex plane, then for $m \ge 4s$ we have the following upper bounds for the error of FPM*

$$\|e_m\|_2 \le C_1 \|f\|_{\Gamma(m)} \left(\frac{s}{m}\right)^{m-1}, \tag{4.6}$$

where

$$C_1 := 24Kes\left(1 + \frac{1}{8s}\right)\|v\|_2. \tag{4.7}$$

Proof. Since $\eta = \infty$, by Theorem 4.4 the upper bound (4.5) is valid for each $s < r < R < +\infty$. Now, if in (4.5) we put $r = s\left(1 + \frac{1}{m}\right)$, $m \geq 1$, we obtain

$$\begin{aligned} s < r \leq 2s, \\ \frac{r+s}{r-s} = 2m+1. \end{aligned}$$

Moreover we have

$$\left(\frac{r}{R}\right)^m = \left(\frac{s}{R}\right)^m \left(1 + \frac{1}{m}\right)^m \leq e\left(\frac{s}{R}\right)^m. \tag{4.8}$$

Now, since R can be chosen arbitrarily large, we can put $R = m$, so that, for $m \geq 4s$ $(m \geq 2r)$,

$$\frac{R+r}{R-r} = \frac{m+r}{m-r} \leq 3, \tag{4.9}$$

$$\frac{1}{1-\left(\frac{r}{R}\right)^m} = \frac{1}{1-\left(\frac{r}{m}\right)^m} \leq \frac{1}{1-\frac{r}{m}} \leq 2, \tag{4.10}$$

$$2m+1 \leq 2m\left(1 + \frac{1}{2m}\right) \leq 2m\left(1 + \frac{1}{8s}\right). \tag{4.11}$$

Inserting (4.8), (4.9), (4.10), (4.11) in (4.5) the theorem is proved. □

THEOREM 4.6. *Let f be analytic in the interior of $\Gamma(\eta)$, $\eta < \infty$. Assume that $W(A) \subseteq \Omega(s)$, $s < \eta$. If $\overline{m}$ is the smallest integer such that*

$$s\left(1 + \frac{1}{\overline{m}}\right) < \eta,$$

then for $m \geq \overline{m}$, for the FPM we have the following upper bounds

$$\|e_m\|_2 \leq C_2 m \|f\|_{\Gamma\left(\eta - \frac{\varepsilon}{m}\right)} \left(\frac{s}{\eta}\right)^m, \tag{4.12}$$

where

$$\varepsilon = \varepsilon(m) = \eta - s\left(1 + \frac{1}{m}\right),$$

and

$$C_2 = 4K\left(1 + \frac{T}{2}\right) \frac{\left(1 + \frac{sQ}{\eta}\right)}{\left(1 - \frac{(\eta - s)T}{\eta} - \frac{sQ}{\eta}\right)} \left(\frac{1}{1 - \frac{(\eta - s)T}{\eta}}\right)^{\overline{m}} \|v\|_2, \tag{4.13}$$

with $Q = 1 + \frac{1}{\overline{m}}$, $T = Q - 1$.

Proof. For $m \geq \overline{m}$ let

$$r = s\left(1 + \frac{1}{m}\right).$$

Thus we have

$$s < r \leq sQ \leq 2s,$$

and

$$\frac{(r+s)}{(r-s)} = 2m+1 = 2m\left(1+\frac{1}{2m}\right) \leq 2m\left(1+\frac{T}{2}\right). \tag{4.14}$$

Now, let us define

$$R = R(m) := \eta - \frac{\varepsilon}{m}, \tag{4.15}$$

where $\varepsilon := \eta - r$. Thus $R \leq \eta$, and $R \to \eta$ for $m \to \infty$. Hence, by (4.15) we obtain

$$\begin{aligned} \frac{R+r}{R-r}\frac{1}{1-\left(\frac{r}{R}\right)^m} &\leq \frac{R\,(R+r)}{(R-r)^2} \\ &\leq \frac{\eta\,(\eta+r)}{\left(\eta-\frac{\varepsilon}{m}-r\right)^2} \\ &\leq \frac{\left(1+\frac{sQ}{\eta}\right)}{\left(1-\frac{(\eta-s)T}{\eta}-\frac{sQ}{\eta}\right)}. \end{aligned} \tag{4.16}$$

Using again (4.15) and the inequalities $\varepsilon \leq \eta - s$, $T \geq (1/m)$, we have

$$\begin{aligned} \left(\frac{s}{R}\right)^m = \left(\frac{s}{\eta}\right)^m \left(\frac{1}{1-\frac{\varepsilon}{\eta m}}\right)^m &\leq \left(\frac{s}{\eta}\right)^m \left(\frac{1}{1-\frac{(\eta-s)}{\eta\overline{m}}}\right)^{\overline{m}} \\ &= \left(\frac{s}{\eta}\right)^m \left(\frac{1}{1-\frac{(\eta-s)T}{\eta}}\right)^{\overline{m}}. \end{aligned} \tag{4.17}$$

Finally, inserting (4.14), (4.16) and (4.17) in (4.5), we get the thesis. □

5. Some applications. In this section we consider some examples of problems involving the computation of a matrix function times a vector. For these examples we want to specialize the error estimates given in previous section for the FPM. In order to deal with practical problems, where necessary we investigate the operation $f(tA)v$, where $t > 0$ instead of $f(A)v$. It is important to point out that in order to get error bounds for the FPM, it is fundamental to locate the singularity of the function f. In particular, if f is analytic in the whole complex plane we can get error bounds using Theorem 4.5, whereas if f has some singularities we use Theorem 4.6. In both cases we have to analyze the quantity

$$\|f\|_{\Gamma(R)},$$

where R must be chosen as in the above cited theorems.

Throughout this section, we assume that $F(A)$ in (4.1) is strictly contained in the right half plane and convex. Moreover, since A is real, we assume to work with compacts symmetric with respect to the real axis. In other words, we shall work with compact subsets belonging to the class

$$\overline{\mathbf{M}} := \left\{ \begin{array}{c} \Omega \in \widetilde{\mathbf{M}} : \Omega \text{ is symmetric with respect to the} \\ \text{real axis, convex and } \Omega \subset \mathbf{C}^+ \end{array} \right\} \tag{5.1}$$

As we shall see, convexity (and symmetry) is a property that allows to simplify some later results, because it implies

$$\min_{z\in\Gamma(R)} \mathrm{Re}(z) = \psi(-R), \tag{5.2}$$

$$\max_{z\in\Gamma(R)} \mathrm{Re}(z) = \psi(R), \tag{5.3}$$

for each $R \geq \gamma$. Moreover, for some results we also assume that Ω has a vertical axis (i.e., Ω is symmetric with respect to a vertical axis), that implies

$$\max_{z\in\Gamma(R)} \mathrm{Im}(z) = \mathrm{Im}\,\psi(iR). \tag{5.4}$$

At least for $R \to \infty$, these assumptions are tolerable. In fact, as proved in [26], there always exists $\widetilde{R}$ such that the compact $\Omega(R)$ is convex for each $R \geq \widetilde{R}$, and so (5.2) and (5.3) hold for $R \geq \widetilde{R}$. Furthermore, as $|w| \to \infty$,

$$\psi(w) = w + \alpha_0 + O\left(1/w\right),$$

that is, $\Gamma(R) \to C(\alpha_0, R)$ as $R \to \infty$, where $C(\alpha_0, R)$ denotes the circle of radius R centered in α_0. Since (5.4) clearly holds for circles, as $R \to \infty$

$$\max_{z\in\Gamma(R)} \mathrm{Im}(z) \to \mathrm{Im}\,\psi(iR).$$

The following result holds also for Ω not convex.

PROPOSITION 5.1. *Let $\Omega \in \widetilde{\mathbf{M}}$, symmetric with respect to the real axis and contained in the right half plane. For each $R \geq \gamma$,*

$$\max_{z\in\Gamma(R)} \mathrm{Re}(z) \leq K_\psi \psi(R),$$

where K_ψ is a constant depending only on ψ.

Proof. Let $\overline{w}$, $|\overline{w}| = R$, be such that

$$\mathrm{Re}(\psi(\overline{w})) = \max_{z\in\Gamma(R)} \mathrm{Re}(z).$$

By symmetry, $\psi(R)$ is real, and so

$$\mathrm{Re}(\psi(\overline{w})) - \psi(R) = \mathrm{Re}(\psi(\overline{w}) - \psi(R)) = \mathrm{Re}\int_R^{\overline{w}} \psi'(u)du.$$

Using the bound

$$\left|\psi'(u)\right| \leq 1 + \left(\frac{\gamma}{|u|}\right)^2, \qquad |w| > \gamma, \tag{5.5}$$

(see [16] for the proof) we get

$$\mathrm{Re}(\psi(\overline{w})) - \psi(R) \leq \int_R^{\overline{w}} \left(1 + \frac{\gamma^2}{R^2}\right) du \leq \pi R.$$

Since $\psi(R) > 0$,

$$\frac{\mathrm{Re}(\psi(\overline{w}))}{\psi(R)} \leq 1 + \pi\frac{R}{\psi(R)}.$$

Finally, using

$$\psi(R) = R\left(1 + \frac{\alpha_0}{R} + \frac{\alpha_1}{R^2} + ...\right)$$

we have the thesis.
□

Similarly we can also prove

$$\min_{z\in\Gamma(R)} \mathrm{Re}(z) \geq \overline{K}_\psi \psi(-R),$$
$$\max_{z\in\Gamma(R)} \mathrm{Im}(z) \leq \widetilde{K}_\psi \,\mathrm{Im}\,\psi(iR),$$

where $\overline{K}_\psi$ and $\widetilde{K}_\psi$ are constants depending on ψ. By all these arguments, we can say that whenever we shall use the hypothesis of convexity or symmetry with respect to a vertical axis with relations (5.2), (5.3) and (5.4), we shall simplify the results without any effective theoretical restrictions.

The following Lemmas will be frequently used.

LEMMA 5.2. *Let $\Omega \in \widetilde{\mathbf{M}}$, symmetric with respect to the real axis, with capacity γ. Given $R > \gamma$, for the associated conformal mapping ψ we have the following relations*

$$\psi(R) \leq \psi(\gamma) + R - \frac{\gamma^2}{R} \tag{5.6}$$
$$\psi(-R) \geq \psi(-\gamma) - R + \frac{\gamma^2}{R} \tag{5.7}$$
$$\mathrm{Im}\left(\psi(iR)\right) \leq \mathrm{Im}\left(\psi(i\gamma)\right) + R - \frac{\gamma^2}{R} \tag{5.8}$$

Proof. Writing

$$\psi(R) = \psi(\gamma) + \int_\gamma^R \psi'(t)dt,$$

where the integral path is the real line segment $[\gamma, R]$, using (5.5) we have

$$\begin{aligned}\psi(R) - \psi(\gamma) = |\psi(R) - \psi(\gamma)| &\leq \int_\gamma^R \left|\psi'(t)\right| dt \\ &\leq \int_\gamma^R \left(1 + \left(\frac{\gamma}{|t|}\right)^2\right) dt. \end{aligned} \tag{5.9}$$

Since t is real, $|t|^2 = t^2$, and we can integrate (5.9) getting (5.6). Analogously, one proves (5.7) and (5.8). □

LEMMA 5.3. *Let $\Omega \in \overline{\mathbf{M}}$. Let s be such that $\gamma \leq s < \phi(0)$. Let $\overline{m}$ be the smallest integer such that*

$$s\left(1 + \frac{1}{\overline{m}}\right) < \phi(0).$$

Defining

$$R = R(m) = \phi(0) - \frac{\varepsilon}{m},$$

where

$$\varepsilon = \varepsilon(m) = \phi(0) - s\left(1 + \frac{1}{m}\right),$$

for $m \geq \overline{m}$ *we have*

$$\psi(-R) \geq \frac{L}{m}, \tag{5.10}$$

where $L > 0$ *is a constant depending on* ψ *and* s.

Proof. In order to simplify the notations, we set $\rho := \phi(0)$ (note that $\phi(0)$ is real because Ω is symmetric with respect to the real axis). Hence, by the definition of R,

$$\begin{aligned}
\psi(-R) &= \psi(-\rho + \frac{\varepsilon}{m}) \\
&= -\rho + \frac{\varepsilon}{m} + \alpha_0 + \frac{\alpha_1}{-\rho + \frac{\varepsilon}{m}} + ... \\
&= -\rho + \frac{\varepsilon}{m} + \alpha_0 - \frac{\alpha_1}{\rho} + \frac{\alpha_1 \frac{\varepsilon}{m}}{\rho\left(-\rho + \frac{\varepsilon}{m}\right)} + ... \\
&= \psi(-\rho) + \frac{\varepsilon}{m}\left(1 + \frac{\alpha_1}{\rho\left(-\rho + \frac{\varepsilon}{m}\right)} + ...\right).
\end{aligned}$$

Thus, since $\psi(-\rho) = 0$ and $\varepsilon = \varepsilon(m) \geq \rho - s\left(1 + \frac{1}{\overline{m}}\right)$ for $m \geq \overline{m}$, there exists a constant $k > 0$ such that

$$\psi(-R) \geq k\frac{\rho - s\left(1 + \frac{1}{\overline{m}}\right)}{m}.$$

Defining

$$L := k\left(\rho - s\left(1 + \frac{1}{\overline{m}}\right)\right), \tag{5.11}$$

we get the thesis. □

5.1. The case of $f(tA) = \exp(-tA)$. Let's consider the computation of

$$y(t) = e^{-tA}v, \qquad t \geq 0. \tag{5.12}$$

As well known (5.12) is the solution of the IVP

$$\begin{cases} Ay + \frac{dy}{dt} = 0, & t > 0, \\ y(0) = v. \end{cases}$$

THEOREM 5.4. *Let us suppose* $\Omega \in \overline{\mathbf{M}}$. *Assume that* $W(A) \subseteq \Omega(s)$, *for some* $s \geq \gamma$. *Then,*

$$\|e_m(t)\|_2 \leq C_1 \exp(t\mu_1)\left(\frac{s\exp(t)}{m}\right)^{m-1}, \qquad m \geq 4s, \tag{5.13}$$

where C_1 *is defined by (4.7) and* $\mu_1 = 1 - \psi(-\gamma)$.

Proof. Let $R = m \geq 4s$ as in Theorem 4.5. The thesis follows easily from Theorem 4.5 using the relation

$$\|\exp(-tz)\|_{\Gamma(m)} = \exp(-t\psi(-m)),$$

that follows by the hypothesis on Ω and (5.7), that implies

$$\exp(-t\psi(-m)) \leq (\exp(t))^{m-1} \exp(t\,(1 - \psi(-\gamma))).$$

□

5.2. The case of $f(A) = A^{-1}$. The computation of $y = A^{-1}v$ is equivalent to solve the linear system $Ay = v$. Clearly, the function $f(z) = 1/z$ is not analytic in the whole complex plane because it is singular at 0.

To evaluate the approximation error, it is necessary to bound $\|f\|_{\Gamma(R)}$, with R as in Theorem 4.6.

THEOREM 5.5. *Let $\Omega \in \overline{\mathbf{M}}$. Assume that $W(A) \subseteq \Omega(s)$ for $\gamma \leq s < \eta = |\phi(0)|$. Let $\overline{m}$ be the smallest integer such that*

$$s\left(1 + \frac{1}{\overline{m}}\right) < \eta.$$

Then, for the error we have the following upper bound

$$\|e_m\|_2 \leq \frac{C_2}{L} m^2 \left(\frac{s}{\eta}\right)^m, \qquad m \geq \overline{m}, \tag{5.14}$$

where C_2 is defined by (4.13) and L by (5.11).

Proof. Defining R as in Theorem 4.6, by the hypothesis on Ω we get

$$\max_{z \in \Gamma(R)} \left|\frac{1}{z}\right| = \frac{1}{\psi(-R)}.$$

Hence, the thesis follows immediately by Theorem 4.6 and Lemma 5.3. □

5.3. The case of $f(A) = \sqrt{A}$. An example of practical application where the computation of the matrix square root arises is the following initial value problem

$$\begin{cases} Au + \frac{d^2u}{dt^2} = 0, \quad t > 0, \\ u(0) = v, \quad \frac{du}{dt}(0) = w, \end{cases} \tag{5.15}$$

whose solution is $u(t) = \cos(t\sqrt{A})v + \sqrt{A^{-1}}\sin(t\sqrt{A})w$.

The function $f(z) = \sqrt{z}$ is not analytic in the whole complex plane because it has a branch point at 0. Here, we want to consider only the branch of the square root such that $\sqrt{1} = 1$. Namely, on the basis of definition (1.2) we set

$$\sqrt{A} = \frac{1}{2\pi i} \int_\Gamma \sqrt{z}(zI - A)^{-1}\, dz.$$

With this assumption, the square root can be considered analytic in all compact subsets not containing 0.

THEOREM 5.6. *Under the same hypothesis of Theorem 5.5 and with the further hypothesis that Ω has a vertical axis, for the error we have the following upper bounds*

$$\|e_m\|_2 \leq C_2 \mu_2 m \left(\frac{s}{\eta}\right)^m, \qquad m \geq \overline{m}, \tag{5.16}$$

where C_2 is defined by (4.13) and

$$\mu_2 := \sqrt{\psi(\eta)^2 + \operatorname{Im}\psi(i\eta)^2}. \tag{5.17}$$

Proof. Let R be as in Theorem 4.6. By the hypothesis on Ω, writing $z = x + iy$,

$$\begin{aligned}\max_{z\in\Gamma(R)} \left|\sqrt{z}\right| &= \max_{z\in\Gamma(R)} \sqrt{|z|} \\ &= \max_{z\in\Gamma(R)} \sqrt{x^2+y^2} \\ &\leq \sqrt{\psi(R)^2 + \operatorname{Im}\psi(iR)^2}\end{aligned}$$

where we used (5.3) and (5.4). Since $R < \eta$, we have simply

$$\max_{z\in\Gamma(R)} \left|\sqrt{z}\right| \leq \sqrt{\psi(\eta)^2 + \operatorname{Im}\psi(i\eta)^2}.$$

Hence, by Theorem 4.6, we get the thesis. □

5.4. The case of $f(A) = \cos(A)$. The computation of the matrix cosine arises in important application such as in the solution of (5.15). Since the cosine function is analytic in the whole complex plane, we have the following.

THEOREM 5.7. *Let $\Omega \in \overline{\mathbf{M}}$, with a vertical axis. Assume that $W(A) \subseteq \Omega(s)$ for $\gamma \leq s$. For the error we have the following upper bound*

$$\|e_m\|_2 \leq C_1 \mu_3 \exp(m) \left(\frac{s}{m}\right)^{m-1}, \qquad m \geq 4s, \tag{5.18}$$

where C_1 is defined by (4.7), and

$$\mu_3 := \frac{1}{2}\left(\left(1 + \frac{1}{\exp(4s)}\right)\cosh\left(\operatorname{Im}\psi(i\gamma)\right) + \sinh\left(\operatorname{Im}\psi(i\gamma)\right)\right).$$

Proof. Let $R = m \geq 4s$ as in Theorem 4.5. In order to estimate $\|\cos\|_{\Gamma(m)}$, writing $z = x + iy$ we have

$$\begin{aligned}\cos z &= \cos x \sin iy - \cos iy \sin x \\ &= \cosh y \sin x - i \cos x \sinh y\end{aligned}$$

and thus

$$\begin{aligned}|\cos z| &= \sqrt{1 - \cos^2(x) + \sinh^2(y)} \\ &\leq \sqrt{1 + \sinh^2(y)} \\ &= \cosh(y)\end{aligned}$$

Using this relation and all the hypothesis on Ω, have

$$\|\cos\|_{\Gamma(m)} \leq \cosh\left(\operatorname{Im}\psi(im)\right).$$

Now, using (5.8),

$$\operatorname{Im}\psi(im) \leq \operatorname{Im}\psi(i\gamma) + m,$$

so that

$$\begin{aligned}
\cosh\left(\operatorname{Im}\psi(im)\right) \leq & \\
& \leq \cosh\left(\operatorname{Im}\psi(i\gamma) + m\right) \\
& = \cosh m \cosh\left(\operatorname{Im}\psi(i\gamma)\right) + \sinh m \sinh\left(\operatorname{Im}\psi(i\gamma)\right) \\
& \leq \left(\frac{\exp(m)}{2} + \frac{1}{2}\right)\cosh\left(\operatorname{Im}\psi(i\gamma)\right) + \frac{\exp(m)}{2}\sinh\left(\operatorname{Im}\psi(i\gamma)\right) \\
& = \frac{\exp(m)}{2}\left(\left(1 + \frac{1}{\exp(m)}\right)\cosh\left(\operatorname{Im}\psi(i\gamma)\right) + \sinh\left(\operatorname{Im}\psi(i\gamma)\right)\right) \\
& \leq \mu_3 \exp(m)
\end{aligned}$$

Hence, the thesis follows directly by Theorem 4.5. □

5.5. The case of $f(tA) = \cos(t\sqrt{A})$. As already said, the square root function is not single valued. In fact it is two valued, having a branch for which $\sqrt{1} = 1$ and another for which $\sqrt{1} = -1$. However, since we are working with the cosine, the composite function is single valued. Note that $u(t) = \cos(t\sqrt{A})v$ is the solution of (5.15) with $w = 0$.

THEOREM 5.8. *Let $\Omega \in \overline{\mathbf{M}}$, with a vertical axis and assume that $W(A) \subseteq \Omega(s)$ for $\gamma \leq s$. Let $\widetilde{m} := \lceil \max(\psi(-\gamma), 4s) \rceil$, where $\lceil\cdot\rceil$ denotes the rounding to the nearest integer towards infinity. For the error we have the following upper bound*

$$\|e_m\|_2 \leq C_1\mu_5 \exp(t\mu_4\sqrt{m})\left(\frac{s}{m}\right)^{m-1}, \qquad m \geq \widetilde{m}, \tag{5.19}$$

where C_1 is defined by (4.7), and

$$\mu_4 := \frac{1}{\sqrt{2}}\sqrt{\sqrt{1 + \left(\frac{\operatorname{Im}\psi(i\gamma)}{\widetilde{m}} + 1\right)^2} + 1},$$

$$\mu_5 := \frac{1}{\sqrt{2}}\left(1 + \frac{1}{\exp(2t\mu_4\sqrt{\widetilde{m}})}\right)^{\frac{1}{2}}.$$

Proof. As in Theorem 4.5 let $R = m$. In order to estimate $\left\|\cos(t\sqrt{\cdot})\right\|_{\Gamma(m)}$, writing $z = x + iy$ and defining

$$a := \sqrt{\frac{1}{2}\sqrt{x^2+y^2} + \frac{1}{2}x}, \quad b := \sqrt{\frac{1}{2}\sqrt{x^2+y^2} - \frac{1}{2}x}, \tag{5.20}$$

we have

$$\begin{aligned}\left|\cos\left(t\sqrt{x+iy}\right)\right| &= \sqrt{\cos^2(ta)\cosh^2(tb)+\sin^2(ta)\sinh^2(tb)}\\ &\le \sqrt{\cosh^2(tb)+\sinh^2(tb)}\\ &= \sqrt{\cosh(2tb)}\\ &\le \left(\frac{\exp(2tb)}{2}+\frac{1}{2}\right)^{\frac{1}{2}} \qquad (5.21)\end{aligned}$$

By (5.20), the function $\exp(2tb)$ goes to $+\infty$ as $x \to -\infty$ or $y \to \infty$, so that, under the hypothesis on Ω, we can put $x := \psi(-m)$ and $y := \operatorname{Im}\psi(im)$, obtaining

$$b = \frac{1}{\sqrt{2}}\sqrt{\sqrt{(\psi(-m))^2+(\operatorname{Im}\psi(im))^2}-\psi(-m)}.$$

Using (5.7) and (5.8),

$$\begin{aligned}b &\le \frac{1}{\sqrt{2}}\sqrt{\sqrt{\left(\psi(-\gamma)-m+\frac{\gamma^2}{m}\right)^2+\left(\operatorname{Im}\psi(i\gamma)+m-\frac{\gamma^2}{m}\right)^2}-\left(\psi(-\gamma)-m+\frac{\gamma^2}{m}\right)}\\ &\le \frac{1}{\sqrt{2}}\sqrt{\sqrt{(\psi(-\gamma)-m)^2+(\operatorname{Im}\psi(i\gamma)+m)^2}-(\psi(-\gamma)-m)}\\ &= \frac{1}{\sqrt{2}}\sqrt{m\sqrt{\left(\frac{\psi(-\gamma)}{m}-1\right)^2+\left(\frac{\operatorname{Im}\psi(i\gamma)}{m}+1\right)^2}-\psi(-\gamma)+m}\\ &= \frac{1}{\sqrt{2}}\sqrt{m\left(\sqrt{\left(\frac{\psi(-\gamma)}{m}-1\right)^2+\left(\frac{\operatorname{Im}\psi(i\gamma)}{m}+1\right)^2}-\frac{\psi(-\gamma)}{m}+1\right)}\\ &\le \frac{\sqrt{m}}{\sqrt{2}}\sqrt{\sqrt{1+\left(\frac{\operatorname{Im}\psi(i\gamma)}{\widetilde{m}}+1\right)^2}+1} = \mu_4\sqrt{m} \qquad (5.22)\end{aligned}$$

where to get (5.22) we used the hypothesis $m \ge \widetilde{m}$. Hence, by (5.21) and (5.22),

$$\begin{aligned}\left\|\cos(t\sqrt{\cdot})\right\|_{\Gamma(m)} &\le \left(\frac{\exp(2t\mu_4\sqrt{m})}{2}+\frac{1}{2}\right)^{\frac{1}{2}}\\ &\le \exp(t\mu_4\sqrt{m})\frac{1}{\sqrt{2}}\left(1+\frac{1}{\exp(2t\mu_4\sqrt{\widetilde{m}})}\right)^{\frac{1}{2}}\end{aligned}$$

Finally, by Theorem 4.5 we get the thesis. □

6. Numerical implementation. In this section we want to deal with the practical implementation of the FPM. As we shall see, this leads to the construction of an *hybrid procedure* based on this method. In particular, up to now we have assumed to work with methods built on a given compact subset Ω containing $\sigma(A)$, with associated conformal mapping ψ. Obviously, by a numerical point of view, both Ω and ψ have to be computed, and the aim of this section is to provide some important numerical details about this problem.

6.1. The approximation of the spectrum. By Theorem 2.4 we observe that the asymptotic behavior of an asymptotically optimal method built on a compact Ω depends on how tightly Ω contains the spectrum $\sigma(A)$. The aim is to approximate as well as possible the smallest connected compact subset Ω_{opt} such that $\sigma(A) \subseteq \Omega_{opt}$. If Ω_{opt} is convex, then it coincides with the convex hull of the spectrum, $co(\sigma(A))$.

In practice, the common way to build $\Omega \approx \Omega_{opt}$ consists of using an eigenvalue method to yield a certain number of estimates for $\sigma(A)$ and then defining Ω as the compact whose boundary is the polygon obtained joining the outermost points of the set of estimates (cf. [26]). Since we consider A real, $\sigma(A)$ is symmetric with respect to the real axis and then we can also consider a polygon of this kind. Nevertheless, we must point out that if the function f is not analytic in the whole complex plane, in some cases it may be necessary to approximate very well $\sigma(A)$. In fact, if $\sigma(A)$ is very closed to a singular point of f, but such singular point is not contained in Ω_{opt}, it could happen that the estimating phase leads to a compact Ω containing such singular point, so determining the failure of the method. For instance, this can happen when A is highly non-normal (see [8] and [18] for a detailed analysis of this problem). In such a case it is necessary to use a very accurate eigenvalue method, such as the method proposed in [18] based on the Arnoldi algorithm to estimate the field of values of A^{-1}. If f is analytic in the entire complex plane, even a not very accurate Ω yields acceptable results (see Theorem 2.4).

6.2. The computation of the mapping ψ. Given $\Omega \in \mathbf{M}$ whose boundary is a polygon, in order to determine the Laurent expansion of the associated conformal mapping ψ, we can proceed using the scheme proposed in [26], based on the resolution of the *parameters problem* relative to the *Schwarz-Christoffel transformation* associated to ψ, for which we refer to [29]. Obviously, only a finite number of coefficients of this expansion can be determined numerically, and so, setting a priori this number, instead of ψ we obtain the finite expansion of a conformal mapping

$$\widetilde{\psi} : \overline{\mathbf{C}} \setminus \{w : |w| \leq \widetilde{\gamma}\} \to \overline{\mathbf{C}} \setminus \widetilde{\Omega},$$

that is an approximation of ψ, such that $\widetilde{\gamma} \approx \gamma$ and $\widetilde{\Omega} \approx \Omega$. In the particular case that we compute the only first two coefficients of the Laurent expansion of ψ, that is α_0 e α_1, we approximate Ω with an elliptical compact.

In our numerical experiments, for the computation of the Laurent coefficients of ψ we used the software `Schwarz-Christoffel Matlab Toolbox`, written by Driscoll in 1995 (see [2]).

6.3. The algorithm. On the basis of what said above, the phases for the numerical implementation of the FPM are the following:

1. using a suitable eigenvalue method, compute a set $\{\lambda_j\}_{j=1,\ldots,s}$ of estimates of $\sigma(A)$;
2. build the compact Ω with a polygonal boundary, obtained joining the outermost values of the set $\{\lambda_j\}_{j=1,\ldots,s}$;
3. evaluate the first p coefficients of the Laurent series expansion of the associated conformal mapping ψ, obtaining a mapping $\widetilde{\psi}$ with a finite expansion relative to a compact $\widetilde{\Omega} \approx \Omega$ (if p is not too small, $\widetilde{\Omega}$ is very closed to Ω, but with corners rounded off (see e.g. [20], [21]));
4. compute the Fejèr points and the Newton coefficients;
5. compute the approximation y_m.

REMARK 6.1. *In practical situations, one often has to face problems of type $f(tA)v$. Using the FPM, there are two possible ways to proceed. The first one consists in working with the problem $f(B)v$, where $B := tA$, and the second one consists in working with the function $g(z) := f(tz)$. From the numerical point of view, there are not substantial differences between these two approaches, because once the conformal mapping $\widetilde{\psi}$ relative to $\widetilde{\Omega}$ has been computed, the conformal mapping $\widetilde{\psi}_t$ relative to $t\widetilde{\Omega}$ is given by $\widetilde{\psi}_t(w) = t\widetilde{\psi}(w/t)$, $w > t\widetilde{\gamma}$.*

In our numerical experiments, the eigenvalue estimating phase is performed by means of the Krylov method based on the Arnoldi and Lanczos algorithms. We shall use the following notations: if s is the number of eigenvalue estimates produced by the Arnoldi or Lanczos method, and p is the number of the computed leading Laurent coefficients of ψ, we call ArnFPM(s,p) and LanFPM(s,p), the hybrid procedures obtained applying the FPM to the Arnoldi and Lanczos estimates.

7. Numerical experiments. Let us consider the differential operator

$$-\Delta + \tau_1 \frac{\partial}{\partial x} + \tau_2 \frac{\partial}{\partial y}, \qquad \tau_1, \tau_2 \in \mathbf{R}, \tag{7.1}$$

where Δ denotes the 3-dimensional Laplacian operator. Discretizing using central differences and Dirichlet boundary conditions on the cube $(0,1)\times(0,1)\times(0,1)$, with uniform meshsize $h = 1/(n+1)$ along each direction, a nonsymmetric matrix $\overline{A}$ of order $N = n^3$ with particular block structure is obtained. It can be represented by means of a sum of Kronecker products as follows,

$$\overline{A} := \frac{1}{h^2}\left\{I_n \otimes (I_n \otimes C_1) + [B \otimes I_n + I_n \otimes C_2] \otimes I_n\right\}, \tag{7.2}$$

where the matrix B of order n is defined as

$$B := \begin{bmatrix} 2 & -1 & & \\ -1 & 2 & -1 & \\ & -1 & \ddots & \ddots \\ & & \ddots & \ddots \end{bmatrix},$$

I_n is the identity of order n, and

$$C_i := \begin{bmatrix} 2 & -1+\tau_i\frac{h}{2} & & \\ -1-\tau_i\frac{h}{2} & 2 & -1+\tau_i\frac{h}{2} & \\ & -1-\tau_i\frac{h}{2} & \ddots & \ddots \\ & & \ddots & \ddots \end{bmatrix} \in \mathbf{R}^{n\times n}, \qquad i = 1,2.$$

Using the test matrix $A = h^2\overline{A}$, with $N = 3375$, $\tau_1 = 0$, $\tau_2 = 80$, in the following picture (Fig.7.1) we can see an example of computation of Fejèr points. In particular, referring to the algorithm of §6.3, in this picture we can see: 1) $s = 30$ eigenvalue estimates $\{\lambda_j\}_{j=1,\ldots,30}$ obtained with 30 iterations of the Arnoldi algorithm; 2) the boundary of $\widetilde{\Omega}$ relative to the conformal mapping $\widetilde{\psi}$ obtained computing the first $p = 6$ leading coefficients of the conformal mapping ψ relative to the polygonal compact obtained joining the outermost points of $\{\lambda_j\}_{j=1,\ldots,30}$; 3) 16 Fejèr points relative to $\widetilde{\psi}$.

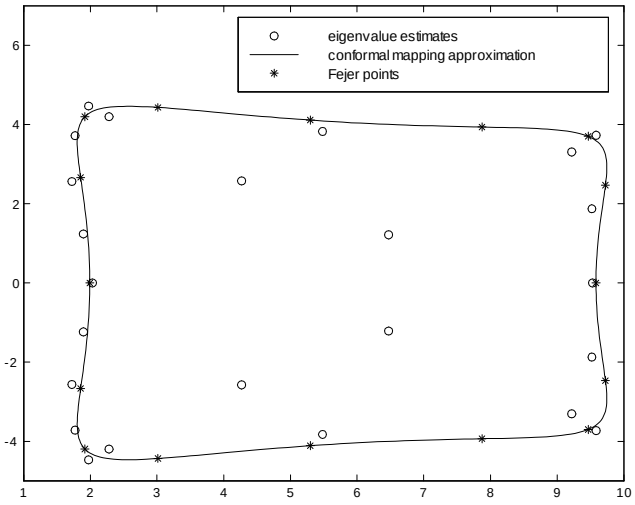

Fig.7.1 $\tau_1 = 0$, $\tau_2 = 80$

Now, using the same matrix $A = h^2\overline{A}$, in the following examples we make a comparison between the hybrid method introduced and the KPMs applied to the computation of the matrix functions treated in Section 5. In all cases we use $v = (1, 1, ..., 1)^T$ and this vector is also used as starting vector for the KPMs. In all pictures, the behavior of $\log_{10} \|e_m\|_2$ with respect to the number of scalar (dot) products is shown. Since A is epthadiagonal, a matrix-vector multiplication costs 7 scalar products. The count of scalar products does not take into account the initial cost of computing the eigenvalues, because in practical situations one usually has to compute more than one or a lot of $f(A)v$ always with the same matrix A but with different vectors v. In all tests we choose $n = 15$, so that the dimension of the problem is $N = 3375$ and $h = 1/16$.

The first consideration we can make observing the pictures below is that the hybrid FPM performs well in each of the cases considered, with a relatively small number of computed eigenvalues.

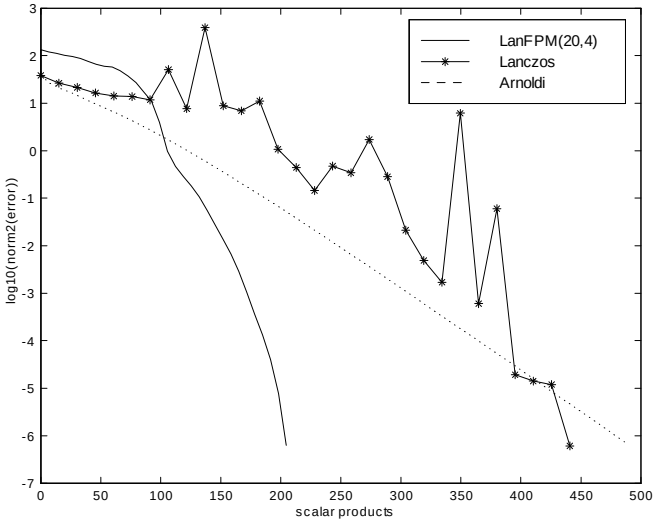

Fig.7.2 $f(A) = e^{-A}$, $\tau_1 = \tau_2 = 70$

By comparing pictures 7.2, 7.5, 7.6 with pictures 7.3 and 7.4, we must also observe that the stability behavior of the hybrid FPM is sensitive to the singularities of f. In other words, as one can expect from the error analysis of Sections 4 and 5, the existence of singularities causes slow down and instability. Nevertheless, as we can see, the regularity of f influences also the performance of the KPMs.

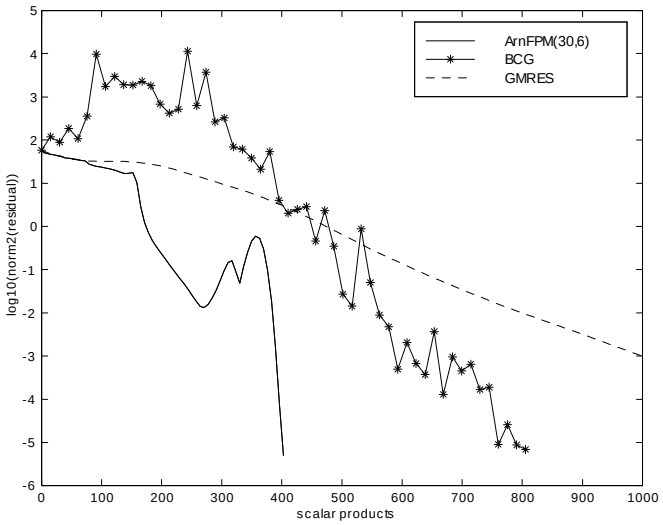

Fig.7.3 $f(A) = A^{-1}$, $\tau_1 = 0$, $\tau_2 = 40$

In Fig. 7.3, since we are solving the linear system $Ay = v$, the KPMs we use as comparison methods are the BCG (that is equivalent to the Lanczos procedures applied to the computation of $A^{-1}v$) and the GMRES (for this example, its restarted version does not show any effective improvement). The results could appear confusing because the GMRES is the optimal polynomial method. Anyway, as well known, such optimality does not concern the cost. For this example the GMRES is faster than the FPM with respect to the number of iterations, but not with respect to the workload.

REMARK 7.1. *As well known, in the case of linear systems, the key issue for Krylov methods is that of preconditioning. A good preconditioner has the effort of collapsing the spectrum of the preconditioned matrix around the point* 1 *of the complex plane. In this way, applying the FPM one gets a compact* Ω *with a small capacity* γ *and such that the quantity* η *(that can be seen as a measure of the distance between* Ω *and the singular point* 0*) is very large with respect to* γ*. Therefore, looking at the asymptotic convergence factor (2.6), we can understand that a good preconditioner can improve remarkably not only Krylov type methods, but also any asymptotically optimal method (see e.g. [10] for some illuminating numerical experiments).*

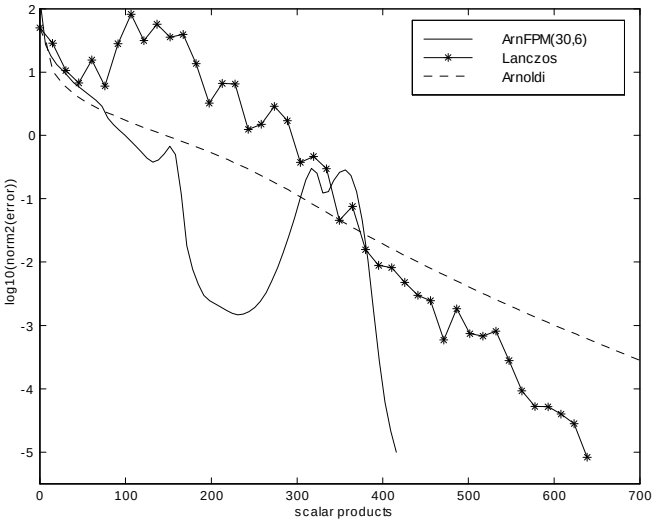

Fig.7.4 $f(A) = \sqrt{A}$, $\tau_1 = 0$, $\tau_2 = 80$

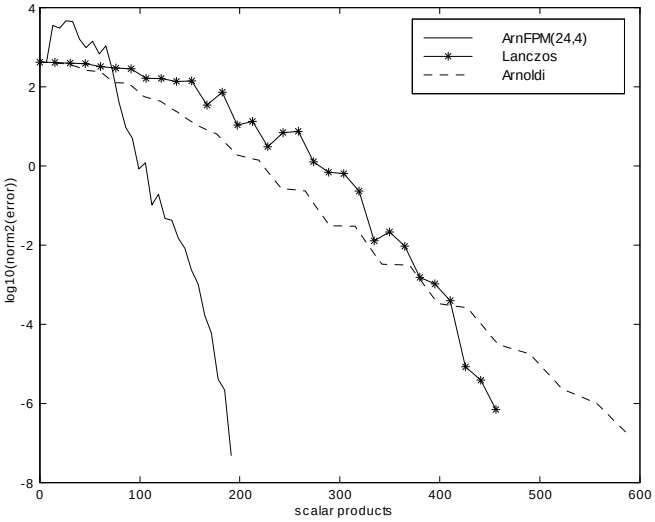

Fig.7.5 $f(A) = \cos(A)$, $\tau_1 = 80$, $\tau_2 = 40$

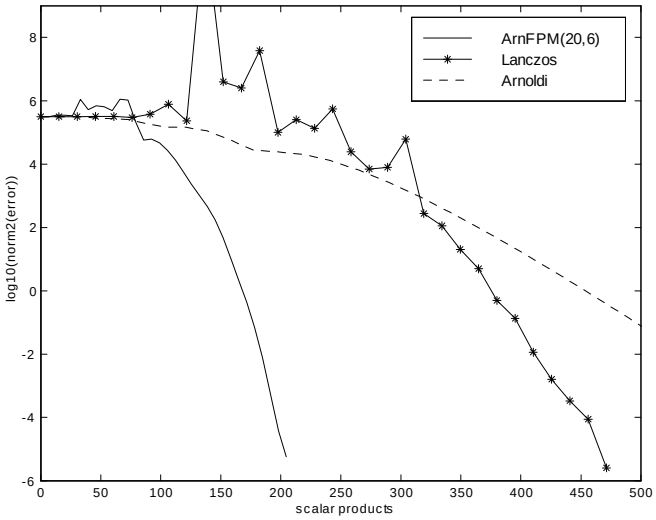

Fig.7.6 $f(A) = \cos(t\sqrt{A})$, $t = 0.5$, $\tau_1 = 10$, $\tau_2 = 100$

Finally, in Fig.7.7 we make a comparison between the FPM and the (truncated) Taylor expansion method in the case of the matrix exponential. The slow down of the Taylor expansion method is due to the fact that it does not use any information on the spectrum and provides an asymptotically optimal approximation with respect to the circles centered at the origin.

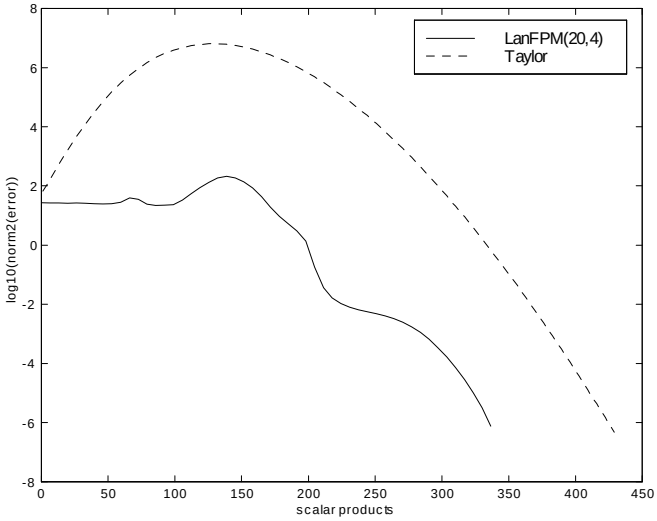

Fig.7.7 $f(A) = e^{-A}$, $\tau_1 = \tau_2 = 170$

REMARK 7.2. *In practical situation, when the exact solution of a certain problem is not known, in order to monitor the convergence of the FPM we have to approximate the constants appearing in the bounds given for* $\|e_m\|_2$, *that is, the bounds given in Theorems 5.4, 5.5, 5.6, 5.7, 5.8. Once the quantity* s *of those formulas has been approximated, the other constants can be easily computed. Nevertheless, in order to compute* s *it is necessary to know* $F(A)$. *In some special cases of* A *it is possible to embrace* $F(A)$ *using some ad-hoc strategies, but in general, one is forced to use the standard theoretical results (see e.g. [14] p.79).*

On the other hand, if one is only required to know how many iterations are nec-

essary to get a given accuracy, i.e., $\|e_m\|_2 < \epsilon$*, it is convenient to use Theorem 2.4. Once the compact* Ω *with capacity* γ *has been computed, under the hypothesis that* $\sigma(A) \subseteq \Omega$*, we have*

$$\overline{\lim_{m\to\infty}} \left[\|e_m\|^{1/m}\right] \leq \frac{\gamma}{\eta},$$

so that one can stop the procedure when $(\gamma/\eta)^m \leq \epsilon$*.*

8. Conclusions. As mentioned in the introduction, the use of Fejèr nodes has never been tested in the context of the computation of functions of matrices. However, because of its low cost and nice theoretical convergence properties, in the opinion of the author it is an efficient method, especially when the function f is analytic in the whole complex plane. If f is not analytic in $\mathbf{C}$, the trend of the error is generally satisfactory but there are some stability problems. A more rigorous analysis of this phenomenon is currently under study.

REFERENCES

[1] L. Bergamaschi and M. Vianello, *Efficient computation of the exponential operator for large, sparse, symmetric matrices*, Numer. Linear Algebra Appl., 7 (2000), pp. 27-45 .

[2] T.A. Driscoll, *Algorithm 756: A MATLAB toolbox for Schwarz-Christoffel mapping*, ACM Trans. Math. Softw., 22 (1996), pp. 168-186.

[3] V. Druskin, A. Greenbaum and L. Knizherman, *Using nonorthogonal Lanczos vectors in the computation of matrix functions*, SIAM J. Sci. Comput., 19 (1998), pp. 38-54.

[4] V. Druskin and L. Knizherman, *Two polynomial methods for calculating functions of symmetric matrices*, U.S.S.R. Comput. Maths. Math. Phys., 29 (1989), pp. 112-121 (English Edition by Pergamon Press).

[5] V. Druskin and L. Knizherman, *Krylov subspace approximations of eigenpairs and matrix functions in exact and computer arithmetic*, Numer.Lin.Alg.Appl., 2 (1995), pp. 205-217.

[6] V. Druskin and L. Knizherman, *Extended Krylov subspaces: approximation of the matrix square root and related functions,* SIAM J. Matrix Anal. Appl., 19 (1998), pp. 755-771.

[7] M. Eiermann, *On semiiterative methods generated by Faber polynomials*, Numer. Math., 56 (1989), pp. 139-156.

[8] M. Eiermann, *Fields of values and iterative methods*, Linear Algebra Appl., 180 (1993), pp. 167-197.

[9] M. Eiermann, W. Niethammer and R.S. Varga, *A study of semiiterative methods for nonsymmetric systems of linear equations,* Numer. Math., 47 (1985), pp. 505-533.

[10] H.C. Elman, Y. Saad, P.E. Saylor, *A hybrid Chebyshev Krylov subspace algorithm for solving nonsymmetric systems of linear equations*, SIAM J. Sci. Stat. Comput., 7 (1986), pp. 840-855.

[11] E. Gallopoulos and Y. Saad, *Efficient solution of parabolic equations by Krylov approximation methods*, SIAM J. Sci. Stat. Comput., 13 (1992), pp. 1236-1264.

[12] F.R. Gantmacher, *The theory of matrices. Vol. 1.* Transl. from the Russian by K. A. Hirsch. Reprint of the 1959 translation. Providence, RI: AMS Chelsea Publishing, 1998.

[13] M. Hochbruck and C. Lubich, *On Krylov subspace approximations to the matrix exponential operator*, SIAM J. Numer. Anal., 34 (1997), pp. 1911-1925.

[14] A.S. Householder, *The theory of matrices in numerical analysis*. New York: Blaisdell, 1964.

[15] L. Knizherman, *Calculation of functions of unsymmetric matrices using Arnoldi's method*, U.S.S.R. Comput. Maths. Math. Phys., 31 (1991), pp. 1-9.

[16] T. Kovari and C. Pommerenke, *On Faber polynomials and Faber expansions*, Math. Z., 99 (1967), pp. 193-206.

[17] T.A. Manteuffel, *The Tchebychev iteration for nonsymmetric linear systems*, Numer. Math., 28 (1977), pp. 307-327.

[18] T.A. Manteuffel and G. Starke, *On hybrid iterative methods for nonsymmetric systems of linear equations*, Numer. Math., 73 (1996), pp. 489-506.

[19] I. Moret and P. Novati, *An interpolatory approximation of the matrix exponential based on Faber polynomials*, J. Comput. App. Math., 131 (2001), pp. 361-380.

[20] I. Moret and P. Novati, *The computation of functions of matrices by truncated Faber series*, Numer. Func. Anal. and Optimiz., 22 (2001), pp. 697-719.

[21] P. Novati, *Polynomial methods for the computation of functions of large unsymmetric matrices*, PhD Thesis, Università degli Studi di Padova, 2000.
[22] Y. Saad, *Analysis of some Krylov subspace approximations to the matrix exponential operator*, SIAM J. Numer. Anal., 29 (1992), pp. 209-228.
[23] R.B. Sidje, *Expokit: A software package for computing matrix exponentials*, ACM Trans. Math. Software, 24 (1998), pp. 130-156.
[24] V.I. Smirnov and N.A. Lebedev, *Functions of a complex variable. Constructive theory*. Translated by Scripta Technica Ltd. London: Iliffe Books Ltd., 1968.
[25] M.N. Spijker, *Numerical ranges and stability estimates*, Appl. Numer. Math., 13 (1993), pp. 241-249.
[26] G. Starke and R.S. Varga, *A hybrid Arnoldi-Faber iterative method for nonsymmetric systems of linear equations*, Numer. Math., 64 (1993), pp. 213-240.
[27] H. Tal-Ezer, *Spectral methods in time for hyperbolic equations*, SIAM J. Numer. Anal., 23 (1986), pp. 11-26.
[28] H. Tal-Ezer, *Spectral methods in time for parabolic problems*, SIAM J. Numer. Anal., 26 (1989), pp. 1-11.
[29] L.N. Trefethen, *Numerical computation of the Schwarz-Christoffel transformation*, SIAM J. Sci. Stat. Comput., 1 (1980), pp. 82-102.
[30] C.F. Van Loan, *The sensitivity of the matrix exponential*, SIAM J. Numer. Anal., 14 (1977), pp. 971-981.
[31] J.L. Walsh, *Interpolation and approximation by rational functions in the complex domain*. American Mathematical Society. Colloquium Publications. 20. Providence, R.I., 1965.
[32] R.C. Ward, *Numerical computation of the matrix exponential with accuracy estimate*, SIAM J. Numer. Anal., 14 (1977), pp. 600-614.