Sastre, J. (2012). Efficient mixed rational and polynomial approximation of matrix functions. Applied Mathematics and Computation. 218(24):11938-11946. doi:10.1016/j.amc.2012.05.064.

Efficient mixed rational and polynomial approximation of matrix functions

J. Sastre[a]

[a]*Instituto de Telecomunicaciones y Aplicaciones Multimedia, Universitat Politècnica de València, Camino de Vera s/n, 46022-Valencia (Spain)*

Abstract

This paper presents an efficient method for computing approximations for general matrix functions based on mixed rational and polynomial approximations. A method to obtain this kind of approximation from rational approximations is given, reaching the highest efficiency when transforming nondiagonal rational approximations with a higher numerator degree than the denominator degree. Then, the proposed mixed rational and polynomial approximation can be successfully applied for matrix functions which have any type of rational approximation, such as Padé, Chebyshev, etc., with maximum efficiency for higher numerator degrees than the denominator degrees. The efficiency of the mixed rational and polynomial approximation is compared with the best existing evaluating schemes for general polynomial and rational approximations, providing greater theoretical accuracy with the same cost in terms of matrix multiplications. It is well known that diagonal rational approximants are generally more accurate than the corresponding nondiagonal rational approximants which have the same computational cost. Using the proposed mixed approximation we show that the above statement is no longer true, and nondiagonal rational approximants are in fact generally more accurate than the corresponding diagonal rational approximants with the same cost.

Keywords: rational approximation, polynomial approximation, mixed rational and polynomial approximation, matrix function
PACS: 87.64.Aa

Email address: jorsasma@iteam.upv.es (J. Sastre)

1. Introduction

Matrix functions play a fundamental role in many areas of science and engineering and are an important subject of study in pure and applied mathematics [1, 2]. Evaluating a function $f(A)$ of an n-by-n matrix A is a frequently occurring problem and many techniques for its computation have been proposed. The main methods for general functions are those based on polynomial approximations, rational approximations, similarity transformations and matrix iterations [1].

In this paper we propose the use of mixed rational and polynomial approximations to compute matrix functions. We show that this kind of approximation can be more efficient than polynomial and rational approximations which provide similar theoretical accuracy. Moreover, we show their relation with rational approximations, and provide a method to obtain the mixed approximations from rational approximations whenever they exist for the considered matrix function, showing that the mixed rational and polynomial approximation can be applied to matrix functions which have any type of rational approximation, such as Padé, Chebyshev, etc., reaching maximum efficiency for nondiagonal rational approximations.

Throughout this paper $\mathbb{R}^{n\times n}$ and $\mathbb{C}^{n\times n}$ denote the sets of real and complex matrices of size $n\times n$, respectively, and I denotes the identity matrix for both sets. $\mathbb{Z}$ denotes the set of integers, $\lceil x \rceil$ denotes the lowest integer not less than x and $\lfloor x \rfloor$ denotes the highest integer not exceeding x.

We will describe the cost of the computations counting the number of matrix operations, denoting by M the cost of a matrix multiplication, and by D the cost of the solution of a multiple right-hand side linear system $AX = B$, where matrices A and B are $n \times n$.

This paper is organized as follows. Section 2 summarizes results for efficient polynomial and rational approximations for general matrix functions. Section 3 deals with the proposed mixed rational and polynomial approximation. Section 4 gives a method to obtain the mixed approximations from nondiagonal rational approximations, and discusses some rounding error issues in the evaluation of the mixed approximation. Section 5 studies the

ional

and polynomial approximations. Finally, conclusions are given in Section 6.

2. Polynomial and rational approximations

Many methods for the computation of matrix functions based on polynomial and rational approximations have been proposed [2, 1]. Among them, the most widely used techniques are those based on Taylor, Padé and Chebyshev approximations. In the following subsections we summarize results for the cost of evaluating polynomial and rational approximations, and some basic properties for Taylor, Padé and Chebyshev approximations.

2.1. Polynomial approximations of matrix functions

Among the different polynomial approximations, Taylor series is a basic tool for computing matrix functions, see Section 4.3 of [1, p. 76-78]. Let $f(A)$ be a matrix function defined by a Taylor series that converges for the square matrix A. Then, we denote $T_m(A)$ the matrix polynomial of degree m that defines the truncated Taylor series of $f(A)$. For $x \in \mathbb{C}$ the truncated Taylor series $T_m(x)$ of a scalar function $f(x)$ about the origin satisfies

$$f(x) - T_m(x) = O(x^{m+1}), \tag{1}$$

and, from now on, we will refer to m as the order of the Taylor approximation.

Below we retrieve some results for the cost of evaluating a matrix polynomial in terms of matrix multiplications M. From (4.3) of [1, p. 74] it follows that the cost of evaluating a matrix polynomial of degree m

$$P_m(A) = \sum_{k=0}^{m} b_k A^k \,, \tag{2}$$

using Horner and Paterson-Stockmeyer's methods [3] is

$$(s + r - 1 - g(s,m))M, \text{ with } r = \lfloor m/s \rfloor, \quad g(s,m) = \begin{cases} 1 & \text{if } s \text{ divides } m, \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

and this quantity is approximately minimized by $s = \sqrt{m}$, so we can take $s = \lceil \sqrt{m} \rceil$ or $s = \lfloor \sqrt{m} \rfloor$, giving both values the same cost [1, p. 74].

From [4, p. 6454-6455], see Table 4.1 of [1, p. 74], using Horner and Paterson-Stockmeyer's methods the maximum degrees of the matrix polynomial (2) that can be evaluated for a given number of matrix products are

$$m \ = \{1, 2, 4, 6, 9, 12, 16, 20, 25, 30, 36, \ldots\} \,, \tag{4}$$

m^*	1	2	4	6	9	12	16	20	25	30
C_P	0	1	2	3	4	5	6	7	8	9

Table 1: Cost C_P in terms of matrix multiplications for the evaluation of polynomial $P_m(A)$ with the first ten values of m .

i.e. for $m = s^2$ (odd positioned elements in m^*) and $m = s(s+1)$ (even positioned elements in m^*), both with $s = 1, 2, 3, \ldots$ The evaluation of matrix polynomial (2) for the degrees in m^* can be performed with minimum cost using evaluation formula (23) of [4, p. 6455]

$$\begin{aligned} P_m(A) = & \left\{ \left\{ \cdots \left\{ b_m A^s + b_{m\ \ 1} A^{s-1} + \ldots + b_{m\ \ s+1} A + b_{m-s} I \right\} \right. \right. \\ & \times A^s + b_{m\ \ s-1} A^{s\ \ 1} + b_{m-s\ \ 2} A^{s\ \ 2} + \ldots + b_{m\ \ 2s+1} A + b_{m-2s} I \Big\} \\ & \times A^s + b_{m\ \ 2s-1} A^{s\ \ 1} + b_{m-2s\ \ 2} A^{s-2} + \ldots + b_{m\ \ 3s+1} A + b_{m-3s} I \Big\} \\ & \cdots \\ & \times A^s + b_{s-1} A^{s-1} + b_{s-2} A^{s-2} + \cdots + b_1 A + I b_0, \end{aligned} \tag{5}$$

after computing and saving matrix powers $A^2, A^3, \ldots, A^s$, where one can take $s = \lceil \sqrt{m} \rceil$ or $s = \lfloor \sqrt{m} \rfloor$. Both selections of s divide m and give the same total cost. Hence, the cost of evaluating (5), denoted by C_P, is $(s \quad 1)M$ to compute the matrix powers, plus $(m/s \quad 1)M$ for evaluating the remaining matrix products in (5), to give

$$C_P = (r + s - 2)M, \text{ with } r = \lfloor m/s \rfloor = m/s, \tag{6}$$

see (3). Table 1 presents the cost C_P of evaluating (5) in terms of matrix multiplications for the first ten values of m^*.

One of the most studied matrix functions is the matrix exponential $\exp(A)$. For $A \in \mathbb{C}^{n \times n}$ the matrix exponential can be defined by the Taylor series

$$\exp(A) = \sum_{k \geq 0} \frac{A^k}{k!}. \tag{7}$$

tions

has been proposed for its computation in [5]. The algorithm uses Horner and Paterson-Stockmeyer's evaluation schemes and is competitive with state-of-the-art algorithms in the literature, such as the one proposed in [6]. New

kinds of polynomial approximations based on series of orthogonal matrix polynomials have been proposed for the matrix exponential and other matrix functions in [4, 7, 8].

2.2. Rational approximations of matrix functions

Rational functions

$$r_{km}(x) = \frac{p_{km}(x)}{q_{km}(x)}, \tag{8}$$

where p_{km} and q_{km} are polynomials in x of degree at most k and m, respectively, are also basic tools for approximation, and they are better able than polynomials to mimic the behavior of general nonlinear functions, see Section 4.4 [1, pp. 78-81]. Two of the main classes of rational approximations used in the computation of matrix functions are Chebyshev and Padé approximations.

We summarize some basic results for rational approximations from [1]. Let $\mathcal{R}_{k,m}$ denote the space of rational functions with numerator and denominator of degrees at most k and m, respectively. The rational function r is a Chebyshev approximation to f on $[a, b]$ from $\mathcal{R}_{k,m}$ if

$$||r(x) - f(x)||_\infty = \min_{s \in \mathcal{R}_{k,m}} ||s(x) - f(x)||_\infty, \tag{9}$$

where $||g||_\infty = \max_{x \in [a,b]} |g(x)|$. Chebyshev approximation is usually only employed for Hermitian matrices, so that error bounds for the scalar problem translate directly into error bounds at a matrix level.

The rational scalar function $r_{km}(x) = p_{km}(x)/q_{km}(x)$ is a $[k/m]$ Padé approximant of scalar function $f(x)$ if $r_{k,m} \in \mathcal{R}_{k,m}$, $q_{km}(0) = 1$, and

$$f(x) - r_{km}(x) = O(x^{k+m+1}). \tag{10}$$

From now on, d_R will denote the degree of the last term of the Taylor series of f about the origin that $r_{km}(x)$ reproduces, i.e. $d_R = k + m$, and we will refer to d_R as the order of the Padé approximation. If a $[k/m]$ Padé approximant exists then it is unique, and it is usually required that p_{km} and q_{km} have no common zeros, so that they are unique. For a given f, k and m, a $[k/m]$ Padé approximant might not exist, though for certain f existence has been proved for all k and m. That is the case of the matrix exponential [9], where using (10.23) [1, p. 241] the $[k/m]$ Padé approximant $r_{km}(A)$ of the matrix exponential is defined by

$$r_{km}(A) = p_{km}(A) \left(q_{km}(A)\right)^{-1}, \tag{11}$$

where

$$p_{km}(A) = \sum_{j=0}^{k} \frac{(k+m-j)!k!}{(k+m)!(k-j)!j!} A^j, \quad q_{km}(A) = \sum_{j=0}^{k} \frac{(-1)^j (k+m-j)!m!}{(k+m)!(m-j)!j!} A^j. \tag{12}$$

Classical theory states that nondiagonal rational approximants, i.e. r_{km} with $k \neq m$, are less accurate than diagonal rational approximants r_{jj}, where $j = \max(k, m)$, and that r_{jj} can be evaluated at a matrix argument with the same cost [9, p. 11], [2, p. 573], [1, p. 242]. In the following sections we show that the last part of this statement is no longer true.

Note that the multiplication by the matrix inverse in (11) is calculated as the solution of a multiple right-hand side linear system. From [10] the cost of the matrix product for $n \times n$ sized matrices and the solution of the multiple right-hand side linear system of the same size are $2n^3 - n^2$ and $\frac{8n^3}{3} - \frac{n^2}{2} + \frac{5n}{6}$ flops, respectively. Therefore, the cost of solving the multiple right-hand side linear system is asymptotically $4/3$ matrix products:

$$D \approx 4/3M. \tag{13}$$

Section 4.4.3 of [1, 80-81] analyzes the cost of evaluating rational functions. When the Paterson-Stockmeyer method is used to evaluate $p_{mm}(A)$ and $q_{mm}(A)$ from a diagonal rational approximation $r_{mm}(A)$, savings can be made: the powers A^2, $A^3, \ldots,$ A^s can be computed once and used in both the evaluations of $p_{mm}(A)$ and $q_{mm}(A)$. From [1, p. 80] the cost of evaluating r_{mm} is then

$$(s + 2r - 1 - 2g(s, m))M + D, \text{ with } r = \lfloor m/s \rfloor, \tag{14}$$

where $g(s, m)$ is defined in (3). This quantity is approximately minimized by $s = \sqrt{2m}$, and therefore one takes for s whichever of $\lfloor \sqrt{2m} \rfloor$ and $\lceil \sqrt{2m} \rceil$ yields the smaller operation count.

From Table 4.2 of [1, p. 80] the maximum values of m that can be obtained for a given number of matrix multiplications in $r_{mm}(A)$ are

$$m^+ = \{1, 2, 3, 4, 6, 8, 10, 12, 15, \ldots\}. \tag{15}$$

Using (13) and (14), the cost of evaluating $r_{mm}(A)$ for the values of m^+ is

$$C_R = (2r + s - 3)M + D \approx (2r + s - 1 \quad 2/3)M, \quad r = m/s, \tag{16}$$

m^+	1	2	3	4	6	8	10	12	15
π_m	0	1	2	3	4	5	6	7	8
C_R	1+1/3	2+1/3	3+1/3	4+1/3	5+1/3	6+1/3	7+1/3	8+1/3	9+1/3
d_R	2	4	6	8	12	16	20	24	30

Table 2: Number of matrix products π_m and cost in terms of matrix multiplications C_R for diagonal rational approximation $r_{mm}(A)$. Approximation order d_R if r_{mm} is a Padé approximant of a given function f.

where s takes whichever value $s = \lceil\sqrt{2m}\rceil$ or $s = \lfloor\sqrt{2m}\rfloor$ divides m and gives the smaller C_R.

From Table 4.2 of [1, p. 80] Table 2 presents the number of matrix products, denoted by π_m, needed for evaluating both $p_{mm}(A)$ and $q_{mm}(A)$ for a diagonal rational function $r_{mm}(A)$ and the first values of m^+, using Horner and Paterson-Stockmeyer's method. This table also presents the total cost C_R of evaluating $r_{mm}(A)$ in terms of matrix products, where we have used (13), and the order d_R of the approximation $r_{mm}(x)$ if it is a Padé approximant of a given function $f(x)$, i.e. $d_R = 2m$.

3. Mixed rational and polynomial approximation.

The proposed method is based on using mixed rational and polynomial approximations of the type

$$z_{jkl}(x) = \frac{u_j(x)}{v_k(x)} + w_l(x), \tag{17}$$

where $u_j(x), v_k(x)$ and $w_l(x)$ are polynomials of x of degrees at most j, k and l, respectively, $u_j(x)$ and $v_k(x)$ have no common zeros, and $v_k(0) = 1$. When x is a matrix in $\mathbb{C}^{n\times n}$, adding polynomial $w_l(x)$ can provide extra accuracy at the same cost in terms of matrix multiplications as the evaluation of the rational expression $u_j(x)v_k^{-1}(x)$. The following example illustrates this fact.

Example 3.1. *From (12) the diagonal Padé approximant $r_{22}(x)$ for the scalar exponential e^x, $x \in \mathbb{C}$, is given by*

$$r_{22}(x) = \frac{x^2/12 + x/2 + 1}{x^2/12 \quad x/2 + 1}, \tag{18}$$

and it satisfies

$$e^x - r_{22}(x) = O(x^5). \tag{19}$$

Thus, the Taylor series of $r_{22}(x)$ reproduces the first five terms of the Taylor series of the exponential e^x about the origin, i.e. $\sum_{j=0}^{4} x^j/j!$. For a matrix A in $\mathbb{C}^{n\times n}$, the computation of Padé approximant $r_{22}(A)$ can be performed efficiently by the evaluation of the matrix power $A_2 = A^2$, and the solution of the corresponding multiple right-hand side linear system to obtain

$$r_{22}(A) = (A_2/12 + A/2 + 1) \times (A_2/12 - A/2 + 1)^{-1}, \tag{20}$$

with total cost $M + D$.

Now, for the matrix exponential we show how to obtain an approximation of the type (17)

$$z_{122}(x) = \frac{u_1(x)}{v_2(x)} + w_2(x), \tag{21}$$

whose cost for square matrices is the same as the cost of evaluating $r_{22}(A)$, and with

$$e^x - z_{122}(x) = O(x^7). \tag{22}$$

Let $u_1(x) = b_1x + b_0$ and $v_2(x) = a_2x^2 + a_1x + 1$. We can obtain the coefficients b_0, b_1, a_1 and a_2 so that Taylor expansion of rational approximation $u_1(x)/v_2(x)$ reproduces the terms of degrees 3, 4, 5, 6 of the exponential Taylor series to give

$$u_1(x) = 55/9x - 125/6, \quad v_2(x) = x^2/30 - x/3 + 1. \tag{23}$$

The first terms of Taylor expansion of rational expression $u_2(x)/v_2(x)$ are

$$-125/6 - 5/6x + 5/12x^2 + x^3/3! + x^4/4! + x^5/5! + x^6/6! + x^7/5400, \ldots \tag{24}$$

Hence, taking

$$w_2(x) = 1/12x^2 + 11/6x + 131/6, \tag{25}$$

and

$$z_{122}(x) = \frac{550x - 1875}{3x^2 - 30x + 90} + \frac{x^2}{12} + \frac{11x}{6} + \frac{131}{6}, \tag{26}$$

where $u_2(x)$ and $v_2(x)$ have been multiplied by 90 so that all the coefficients in the rational part of $z_{122}(x)$ are integer, it follows that $z_{122}(x)$ satisfies (22). For a matrix $A \in \mathbb{C}^{n\ n}$, the evaluation cost of $z_{122}(A)$ is the same as that

for $r_{22}(A)$, i.e. $M + D$. Note that manipulating adequately $z_{122}(x)$ we can obtain an equivalent expression $z_{222}(x)$:

$$z_{122}(x) = z_{222}(x) = \frac{u_2(x)}{v_2(x)} + w_2(x) = \frac{131x^2 - 210x + 180}{6x^2 - 60x + 180} + \frac{x^2}{12} + \frac{11x}{6}, \quad (27)$$

where $u_2(0)/v_2(0) = 1$ and Taylor expansion of the rational part $u_2(x)/v_2(x)$ reproduces the terms of degrees 0, 3, 4, 5 and 6 of the exponential Taylor series.

It is important to note that the Taylor approximation of the same order as $z_{122}(A)$ and $z_{222}(A)$, i.e order 6, can be evaluated efficiently using Horner and Paterson-Stockmeyer's expression (5) in the following way

$$T_6(x) = \sum_{k=0}^{6} \frac{x^k}{k!} = ((A^2/6! + A/5! + I/4!)A^2 + A/3! + I/2)A^2 + A + I, \quad (28)$$

at cost $3M$. Thus, using approximation (26) or (27) with cost $M + D$ instead of (28), we are changing two matrix products for the solution of one multiple right-hand side linear system to obtain the same approximation order. Hence, using (13), for square matrices and the same approximation order, the mixed polynomial and rational approximation saves 2/3 matrix products with respect to the truncated Taylor series. And with respect to Padé diagonal approximation, the mixed polynomial and rational approximation increases the degree of the approximation error order from $O(x^5)$ to $O(x^7)$ at the same cost in terms of matrix multiplications.

Moreover, the following example shows how we can perform aggregations of expressions of type (17) to increase the approximation order with an even lower cost.

Example 3.2. *From (10) the diagonal Padé approximant $r_{55}(x)$ for the scalar exponential e^x, $x \in \mathbb{C}$ satisfies*

$$e^x - r_{55}(x) = O(x^{11}). \quad (29)$$

Using (12), for a matrix A in $\mathbb{C}^{r\times r}$, the Padé approximant $r_{55}(A)$ can be computed by the solution of the corresponding multiple right-hand side system to obtain

$$\begin{aligned} r_{55}(A) &= (A^5/30240 + A^4/1008 + A^3/72 + A^2/9 + A/2 + 1) \\ &\quad \times(-A^5/30240 + A^4/1008 - A^3/72 + A^2/9 - A/2 + 1)\ ^1. \quad (30) \end{aligned}$$

From Table 4.2 of [1, p. 80], the cost of computing $r_{55}(A)$ is 4 M to obtain the numerator and the denominator of $r_{55}(A)$ and the cost of solving the system, to give $4M + D \approx (5 + 1/3)M$ where (13) has been used. However, [11, p. 1183-1184] shows that for the matrix exponential we can exploit the similarity between numerator and denominator coefficients to reduce the work, and using (13) from Table 2.2 [11, p. 1184] the reduced cost is $3M + D \approx (4 + 1/3)M$.

Proceeding as in Example 3.1 we can obtain an approximation of type (17)

$$\begin{aligned} z_{342}(x) &= \frac{u_3(x)}{v_4(x)} + w_2(x) \qquad (31)\\ &= \frac{17696x^3 \quad 223440x^2 + 1651104x \quad 4304160}{15x^4 - 420x^3 + 5040x^2 - 30240x + 75600} + \frac{x^2}{30} + \frac{29x}{15} + \frac{869}{15}, \end{aligned}$$

where $u_3(x)/v_4(x)$ reproduces the terms 3, 4, ..., 10 of the exponential Taylor series, and $z_{342}(x)$ satisfies

$$e^x - z_{342}(x) = O(x^{11}). \qquad (32)$$

As in Example 3.1, we can also obtain an equivalent expression

$$z_{442}(x) = \frac{869x^4 - 6636x^3 + 68544x^2 - 100800x + 75600}{15x^4 \quad 420x^3 + 5040x^2 \quad 30240x + 75600} + \frac{x^2}{30} + \frac{29x}{15} = z_{342}(x), \qquad (33)$$

where the rational part reproduces the exponential Taylor series terms of degrees 0, 3, 4, ..., 10. The evaluation cost of both (31) and (33) for a square matrix A is $3M$ to obtain the powers A^2, A^3, A^4, and D, i.e. $3M + D \approx (4+1/3)M$. The same cost holds if we compute only A^2 and use Horner and Paterson-Stockmeyer's method to compute the numerator and denominator of $z_{442}(A)$ or $z_{342}(A)$.

Now, we proceed to transform (33) into the following expression

$$z_{442}(x) = \frac{u_4(x)}{v_4(x)} + w_2(x) = \left(\frac{u_2^{(2)}(x)}{v_2^{(2)}(x)} + u_1^{(1)}(x) \quad / v_2^{(1)}(x) + w_2'(x), \qquad (34) \right.$$

where $w_2'(x) = x^2/30 + 29x/15$ and

$$v_4(x) = v_2^{(2)}(x) v_2^{(1)}(x), \qquad u_4(x) = u_1^{(2)}(x) + v_2^{(2)}(x) u_1^{(1)}(x). \qquad (35)$$

There are different options for factorizing $v_4(x)$ *into the product* $v_2^{(2)}(x)v_2^{(1)}(x)$. *If real matrices are considered, factorizations with all real coefficients avoid complex arithmetic when evaluating the mixed rational and polynomial approximation* z_{jkl}. *Then, taking into account that, for instance,* $75600 = 280 \times 270$, *we used Matlab's Symbolic Math Toolbox and high precision arithmetic to obtain the following factorization of* $v_4(x)$ *with real coefficients*

$$v_4(x) = (a_2^{(2)}x^2 + a_1^{(2)}x + 280)(a_2^{(1)}x^2 + a_1^{(1)}x + 270), \tag{36}$$

solving the corresponding system of equations, where the values of the constants with 20 significant digits are

$$\begin{aligned} a_2^{(2)} &= 3.6072579301677768986,\ a_1^{(2)} = \quad 44.739103822600446097, \\ a_2^{(1)} &= 4.1582831863931438203,\ a_1^{(1)} = -64.858721313920998406. \end{aligned} \tag{37}$$

This factorization of $v_4(x)$ *gives* $u_2^{(2)}(x) = b_2^{(2)}x^2 + b_1^{(2)}x + b_0^{(2)}$ *and* $u_2^{(1)}(x) = b_2^{(1)}x^2 + b_1^{(1)}x$, *where*

$$\begin{aligned} &b_2^{(2)} = 52459.797022614745322,\ b_1^{(2)} = -422291.34163282687469,\ b_0^{(2)} = 75600, \\ &b_2^{(1)} = 240.90320593170946532,\ b_1^{(1)} = 1148.1833629743816953. \end{aligned} \tag{38}$$

Note that from (36) it follows that $v_4(0) \neq 1$, *as we multiplied the numerator and denominator of the rational part of* $z_{342}(x)$ *by 75600 so that both of them had integer coefficients.*

Using (34), approximation $z_{442}(A)$ *can be evaluated in one matrix product to obtain* $A_2 = A^2$ *and two solutions of multiple right-hand side systems, resulting in* $M + 2D \approx (3 + 2/3)M$. *Thus, (34) can be evaluated saving* $2/3$ *matrix products with respect to the most efficient existing evaluation scheme for matrix exponential rational approximation* $r_{55}(A)$ *[11].*

Taking into account Example 3.2, we propose the following general expression for the aggregation of mixed rational and polynomial approximations

$$\begin{aligned} t_{ijs}(x) = \big(\big(\cdots\big(u_s^{(i)}(x)/v_s^{(i)}(x) + u_s^{(i-1)}(x)\big)/v_s^{(i\ \ 1)}(x) + u_s^{(i-2)}(x)\big) \\ /v_s^{(i\ \ 2)}(x) + u_s^{(i\ \ 3)}(x)\big)/\cdots + u_s^{(1)}(x)\big)/v_s^{(1)}(x) + w_{js}(x), \end{aligned} \tag{39}$$

where $v_s^{(k)}(x)$, $u_s^{(k)}(x)$, $k = 1, 2, \ldots, i$, are polynomials of x with degrees at most s, $w_{js}(x)$ is a polynomial of x with degree at most js, and $i \quad 1$,

$s \geq 0$ and $j \geq 0$. We will show in Section 5 that the definition of w_{js} as a polynomial of degree js instead of just s provides the lowest cost for some orders of approximation.

Note that the number of divisions in (39) is directly equal to variable i. Hence, for convenience, for $i = 0$ we take

$$t_{ijs}(x) = w_{js}(x), \quad i = 0. \tag{40}$$

and then $t_{0\,js}(x)$ is a polynomial.

The following section provides a method to obtain t_{ijs} from rational approximations.

4. Method for obtaining mixed rational and polynomial approximations. Rounding error issues.

Note that using (39) it follows that

$$\begin{aligned} t_{ijs}(x) = & \left(u_s^{(i)}(x) + v_s^{(i)}(x)u_s^{(i\ \ 1)}(x) + v_s^{(i)}(x)v_s^{(i-1)}(x)u_s^{(i-2)}(x)\right. \\ & + \cdots + v_s^{(i)}(x)v_s^{(i-1)}(x)\cdots v_s^{(2)}(x)u_s^{(1)}(x) \\ & \left. + v_s^{(i)}(x)v_s^{(i-1)}(x)\cdots v_s^{(1)}(x)w_{js}(x)\right) / (v_s^{(i)}(x)v_s^{(i-1)}(x)\cdots v_s^{(1)}(x)), \end{aligned} \tag{41}$$

which is a nondiagonal rational expression with numerator of degree at most $(i + j)s$ and denominator of degree is. Hence, if for a given function f there exists any type of nondiagonal rational approximation, such as Padé, Chebyshev, etc., with numerator and denominator degrees $(i + j)s$ and is, respectively, with $i \geq 1$, $s \quad 0$ and $j \geqslant 0$, then we can obtain an equivalent approximation of type (39) by factorizing its denominator in the polynomials $v_s^{(k)}(x)$, $k = 1, 2, \ldots, i$, and then obtaining the polynomials $u_s^{(k)}(x)$, $k = 1, 2, \ldots, i$ and $w_{js}(x)$ by successive polynomial division. For the case of Examples 3.1 and 3.2 it is easy to show that approximations (27) and (33) are equivalent to exponential nondiagonal Padé approximants $r_{42}(x)$ and $r_{64}(x)$, respectively. Analogously, nondiagonal rational Chebyshev approximations can be transformed into equivalent approximations of type (39). For the case where the original nondiagonal rational expression has all real coefficients and s is even, it is possible to obtain t_{ijs} with all real coefficients, avoiding complex arithmetic if $A \in \mathbb{R}^{n\ \ n}$.

Note that we can also obtain approximations of type (39) from diagonal rational approximations. In that case $j = 0$, and therefore w_{js} is null or a

constant. However, Section 5 shows that these approximations are not so efficient as those with $j > 0$.

Now we present an example where we obtain a mixed rational and polynomial approximation from a rational approximation of the matrix cosine:

Example 4.1. *From [1, p. 290] the diagonal Padé approximant $r_{44}(x)$ for the scalar cosine* $\cos(x)$, $x \in \mathbb{C}$, *is given by*

$$r_{44}(x) = \frac{313/15120x^4 - 115/252x^2 + 1}{13/15120x^4 + 11/252x^2 + 1}, \tag{42}$$

and it satisfies

$$\cos(x) - r_{44}(x) = O(x^{10}). \tag{43}$$

From Table 12.1 of [1, p. 290], the computation of Padé approximant $r_{44}(A)$ for a matrix A in $\mathbb{C}^{n\times n}$ can be performed efficiently with total cost $2M + D$. Using MATLAB Symbolic Toolbox the rational Padé approximation $r_{84}(x)$ is given by

$$r_{84}(x) = \frac{\frac{127\,x^8}{22619520} - \frac{x^6}{1360} + \frac{1121\,x^4}{33660} - \frac{271\,x^2}{561} + 1}{\frac{7\,x^4}{67320} + \frac{19\,x^2}{1122} + 1}, \tag{44}$$

Dividing the polynomials of numerator and denominator we obtain

$$t_{114}(x) = \frac{\frac{673621\,x^4}{2716560} + \frac{11839475\,x^2}{769692} + 1}{\frac{7\,x^4}{67320} + \frac{19\,x^2}{1122} + 1} - \frac{21767\,x^2}{1372} + \frac{127\,x^4}{2352}, \tag{45}$$

where the rational part results 1 for $x = 0$. The evaluation cost of $t_{114}(A)$ is the same as that for $r_{44}(A)$, i.e. $2M + D$. Howewer its approximation order is greater

$$\cos(x) - t_{114}(x) = O(x^{14}). \tag{46}$$

Some considerations must be made about the evaluation of the approximation $t_{ijs}(A$
when computing the mixed rational and polynomial approximation t_{ijs} of a function f for a given square matrix A, it is important to verify that matrices $v_s^{(1)}(A)$, $v_s^{(2)}(A), \ldots, v_s^{(i)}(A)$ are nonsingular, and, for an accurate computation, that all of them are well conditioned. The rounding error analysis of each matrix polynomial to be computed in $t_{ijs}(A)$, i.e. $u_s^{(k)}(A)$, $v_s^{(k)}(A)$, $k = 1, 2, \ldots, i$, and $w_{js}(A)$, is given by Theorem 4.5 of [1, p. 74], where the

rounding error in the polynomial coefficients must be taken into account. All the coefficients of t_{ijs} should be obtained to full precision to reduce rounding error in the evaluation of (39). This can be performed using symbolic tools and high precision arithmetic as the task is performed only once for a given function. The denominator roots should be obtained with high precision arithmetic, and the operations performed with them to obtain the final factorizations should be also performed with high precision arithmetic. For an accurate evaluation of polynomial coefficients from the polynomial roots the ideas from [13] can be applied, combined with symbolic tools and high precision arithmetic.

In order to verify the accuracy of the IEEE double precision arithmetic version of the coefficients from (37) and (38) in Example 3.2 we used high precision arithmetic to calculate the coefficients of $v_4(x)$ and $u_4(x)$ from (35) with the values of the double precision coefficients. The relative error for all the coefficients in $v_4(x)$ and $u_4(x)$ was zero, except for the denominator coefficient 15, which had 1.18×10^{-16} relative error, and the numerator coefficients 100800 and 6636, which had 2.89×10^{-16} and 1.37×10^{-16} relative errors, respectively. Note that the three error values were very near the unit roundoff in IEEE double precision arithmetic, i.e. $u = 2^{-53} \approx 1.11 \times 10^{-16}$.

Finally, note that different selections and orderings of the denominator factors $v_s^{(k)}$ produce different polynomials $u_s^{(k)}$, $k = 1, 2, \ldots, i$ and different values of the rounding error. A different selection of polynomial w_{js} also produces different polynomials $u_s^{(k)}$, $k = 1, 2, \ldots, i$, see for example w_2 from (31) and w_2' from (34) and the corresponding rational expression numerators. Suitable factorizations to reduce the rounding error when evaluating $t_{ijs}(A)$ should be used, see [12].

5. Cost analysis

Matrix polynomial $w_{js}(A)$, $j > 0$, can be evaluated efficiently obtaining matrix powers A^2, $A^3, \ldots, A^s$, and then using (5) from Horner and Paterson-Stockmeyer's method, with a total cost given by (6), where $r = \lfloor js/s \rfloor = j$. Taking into account that matrix powers A^2, $A^3, \ldots, A^s$ can be reused to compute the remaining matrix polynomials in (39), the only extra cost of computing (39) for a square matrix A consists of i solutions of multiple right-hand side linear systems. Hence, the total cost for computing (39) is

$$C_{RP} = (s + j - 2)M + iD \approx (s + j - 2 + 4i/3)M, \quad j > 0,\ s > 0,\ i \quad 0 \quad (47)$$

where C_{RP} denotes the mixed rational and polynomial approximation cost in terms of matrix multiplications. If $j = 0$ and w_{js} is null or a constant then it is easy to show that the corresponding cost is

$$C_{RP} = (s-1)M + iD \approx (s-1+4i/3)M, \quad j = 0,\ s > 0,\ i > 0. \tag{48}$$

Note that for the case where approximation (39) is intended to reproduce the first terms of the Taylor series of a given function f, from the results in Section 4 it is equivalent to a $[(i+j)s/is]$ Padé approximant, and then, whenever it exists, t_{ijs} satisfies

$$f(x) - t_{ijs}(x) = O(x^{(2i+j)s+1}). \tag{49}$$

In that case we denote by d_{RP} the order of the mixed rational and polynomial approximation

$$d_{RP} = (2i+j)s. \tag{50}$$

Using (47) and (48), Table 3 presents for $t_{ijs}(x)$ the values i, j, s, the number of matrix products $\tilde{\pi}_{ijs}$, the approximation order d_{RP} if $t_{ijs}(x)$ reproduces the first d_{RP} terms of the Taylor series of a given function f, and the cost C_{RP} in terms of matrix products for the values of i, j, s that maximize d_{RP} for a given cost. Note that in Table 3 $t_{ijs}(A)$ is a matrix polynomial for $d_{RP} = 1$, 2 and 4 because $i = 0$. We have verified that for $d_{RP} = 2, 12$ and 30 there are other combinations of i, j and s that provide the same cost as that shown in Table 3, i.e. $i = 0, 1, 2$, $j = 2, 2, 2$ and $s = 1, 3, 5$, respectively. Note also that for $d_{RP} = 4$ and 16 the values of j are greater than 1. This justifies the selection of w_{js} in (39) as a polynomial of degree that can be greater than s.

In general, if we use Horner and Paterson-Stockmeyer's method to compute a diagonal rational approximation $r_{mm}(A)$ using matrices $A, A^2, \ldots, A^s$, we can evaluate an approximation of type (17)

$$z_{mms}(x) = \frac{u_m(x)}{v_m(x)} + w_s(x), \tag{51}$$

where $u_m(x), v_m(x)$ and $w_s(x)$ are polynomials of x with degrees at most m, m and s, respectively, at the same cost as $r_{mm}(A)$. We have shown that by adding a suitable $w_s(A)$ then $z_{mms}(x)$ can be equivalent to a nondiagonal rational approximation with higher numerator degree than $r_{mm}(A)$, whenever

d_{RP}	1	2	3	4	6	9	10	12	15	16
i	0	0	1	0	1	1	2	1	2	1
j	1	1	1	2	1	1	1	1	1	2
s	1	2	1	2	2	3	2	4	3	4
$\tilde{\pi}_{ijs}$	0	1	0	2	1	2	1	3	2	4
C_{RP}	0	1	1+1/3	2	2+1/3	3+1/3	3+2/3	4+1/3	4+2/3	5+1/3
d_{RP}	20	21	25	28	30	35	36	42	45	49
i	2	3	2	3	2	3	4	3	4	3
j	1	1	1	1	1	1	1	1	1	1
s	4	3	5	4	6	5	4	6	5	7
$\tilde{\pi}_{ijs}$	3	2	4	3	5	4	3	5	4	6
C_{RP}	5+2/3	6	6+2/3	7	7+2/3	8	8+1/3	9	9+1/3	10

Table 3: Number of matrix products $\tilde{\pi}_{ijs}$, approximation order d_{RP} if $t_{ijs}(x)$ reproduces the d_{RP} first terms of the Taylor series of a given function f, and cost in terms of matrix products C_{RP} for the mixed rational and polynomial approximation $t_{ijs}(A)$ and the optimal (i.e. with minimal cost) values of i, j and s.

d_R	2	4	6	8	12	16	20	24	30
d_{RP}	3	6	9	12	16	21	28	36	45

Table 4: Approximation order d_{RP} that can be obtained with the mixed rational and polynomial approximation with the same cost (or lower if there is no equal optimal cost) as diagonal Padé rational approximation with approximation order d_R.

it exists. Furthermore, using aggregations of the mixed rational and polynomial approximations in the form of (39) this advantage can be even greater. Using Tables 2 and 3, Table 4 shows that for the same or even lower cost the mixed rational and polynomial approximation reaches higher approximation orders than diagonal Padé approximations for all values of approximation order d_R, whenever both approximations exist for a given function f. Similarly, nondiagonal rational Chebyshev approximations with higher numerator degree than the denominator degree can be computed at the same cost as the corresponding diagonal rational Chebyshev approximations.

Next we show that given a Taylor approximation about the origin of a matrix function $f(A)$ with order $m \in m$ (see (4)) it is possible to build a

m	1	2	4	6	9	12	16	20	25	30
d_{RP}	1	2	4	6	10	15	21	28	35	42

Table 5: Approximation order d_{RP} that can be obtained with the mixed rational and polynomial approximation with the same cost (or lower if there is no equal optimal cost) as Taylor approximation of order m.

mixed rational and polynomial approximation with the same approximation order and lower cost in the majority of cases, whenever the equivalent non-diagonal Padé approximation exists. The cost of Taylor approximation with order $m \in m^*$ is given by (6). Taking the same order for the mixed rational and polynomial approximation, i.e. $d_{RP} = m$, and the same value of s as in Taylor approximation, from (50) and (6) it follows that

$$i = \frac{r-j}{2}. \tag{52}$$

If $r = m/s$ is odd, taking $j = 1$ from (52), (47) and (6) it follows that the difference between the Taylor cost and the mixed rational and polynomial approximation cost is

$$C_R - C_{RP} = \frac{r-1}{3}. \tag{53}$$

Hence, whenever the $[(i+j)s/is]$ Padé approximation exists for function $f(A)$, with $j = 1$, $i = (r-1)/2$, the cost of the equivalent mixed rational and polynomial approximation is lower than the cost of the Taylor approximation for $r = m/s > 1$. As r was odd, the last condition is accomplished for $m \geq 3s$.

If $r = m/s$ is even, taking $j = 2$ and proceeding analogously it follows that

$$C_R - C_{RP} = \frac{r-2}{3}, \tag{54}$$

and whenever the corresponding $[(i+j)s/is]$ Padé approximation exists for function $f(A)$, with $j = 2$, $i = (r-2)/2$, C_{RP} is lower than C_R for $r = m/s > 2$. As r was even, the last condition is accomplished for $m \quad 4s$.

Using Tables 1 and 3, Table 5 shows that for the same or even lower cost the mixed rational and polynomial approximation reaches higher approximation orders than the Taylor approximation for $m > 6$, whenever both approximations exist for a given function $f(A)$.

Comparing Tables 1, 2 and 3 in general the approximation $t_{ijs}(A)$ is more efficient than both Taylor and Padé approximations for the same approximation order. For instance, for order 30 the mixed rational and polynomial approximation requires $(7 + 2/3)M$, saving $(1 + 1/3)M$, i.e. 14.81%, with respect to Taylor approximation and $(1 + 2/3)M$, i.e. 17.86% with respect to diagonal Padé approximation, and the absolute difference increases with the approximation order. Another interesting property of $t_{ijs}(A)$ is that its computing cost when increasing approximation order grows many times by steps lower than $1M$.

Finally, it is important to note that approximations with other approximation orders than those listed in Table 3 can be obtained with expressions similar to (39) by reducing the degree of one or several of the polynomials involved, for instance by reducing the degree of $w_{js}(x)$. However, such approximations are not optimal because they do not provide maximum approximation order with minimal cost.

6. Conclusions

This paper proposes an approximation for matrix functions based on the aggregation of mixed rational and polynomial approximations. The cost analysis of the new approximation yields that, in general, it is more efficient than polynomial and diagonal rational approximations. We show how to obtain the new kind of approximation from rational approximations whenever they exist for a given function $f(A)$, and, with the use of the mixed rational and polynomial approximation, we can state that nondiagonal rational approximations are in general more efficient than diagonal rational approximations for similar accuracy.

We are currently applying the mixed rational and polynomial approximations proposed here to the algorithm in [5] to compute the matrix exponential, and to other algorithms to compute other transcendental matrix functions.

7. Acknowledgements

This work has been supported by grant PAID-06-011-2020.

References

[1] N. J. Higham, Functions of Matrices: Theory and Computation, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2008.

[2] G. H. Golub, C. V. Loan, Matrix Computations, 3rd Ed., Johns Hopkins Studies in Math. Sci., The Johns Hopkins University Press, 1996.

[3] M. S. Paterson, L. J. Stockmeyer, On the number of nonscalar multiplications necessary to evaluate polynomials, SIAM J. Comput. 2(1) (1973) 60–66.

[4] J. Sastre, J. Ibáñez, E. Defez and P. Ruiz, Efficient orthogonal matrix polynomial based method for computing matrix exponential, Appl. Math. Comput., 217(14) (2011) 6451–6463.

[5] J. Sastre, J. Ibáñez, E. Defez and P. Ruiz, Accurate matrix exponential computation to solve coupled differential models in engineering, Math. Comput. Model., 54 (2011) 1835-1840.

[6] A. H. Al-Mohy, N. J. Higham, A new scaling and squaring algorithm for the matrix exponential, SIAM J. Matrix Anal. Appl. 31(3) (2009) 970–989.

[7] E. Defez, J. Sastre, J. Ibáñez and P. Ruiz, Computing matrix functions solving coupled differential models, Math. Comput. Model., 50(5-6) (2009) 831–839.

[8] E. Defez, L. Jódar, Some applications of Hermite matrix polynomials series expansions, J. Comput. Appl. Math. 99 (1998) 105–117.

[9] C. B. Moler, C. V. Loan, Nineteen dubious ways to compute the expo- 3–49.

[10] S. Blackford, J. Dongarra, Installation guide for LAPACK, LAPACK Working Note 411, Department of Computer Science University of Tenessee, 1999.

[11] N. J. Higham, The scaling and squaring method for the matrix exponential revisited, SIAM J. Matrix Anal. Appl. 26(4) (2005) 1179–1193.

[12] D. Calvetti., E. Gallopoulos and L. Reichel, Incomplete Partial Fractions for Parallel Evaluation of Rational Matrix Functions, J. Comput. Appl. Math., 59 (1995) 349-380.

s,

Numerical Algorithms, 33 (2003) 153-161.