Convergence analysis of a Padé family of iterations for the matrix sector function

Dmitrii B. Karp†, Minghua Lin‡∗

Abstract The main purpose of this paper is to give a solution to a conjecture concerning a Padé family of iterations for the matrix sector function that was recently raised by B. Laszkiewicz et al in [A Padé family of iterations for the matrix sector function and the matrix $p$th root, Numer. Linear Algebra Appl. 2009; 16:951-970]. Using a sharpened version Schwarz’s lemma, we also demonstrate a strengthening of the conjecture.

Keywords: matrix sector function; Padé approximation; rational matrix iteration; hypergeometric identity

2000 AMS subject classifications 65F30; 33C05

1. Introduction

Let $p \geq 2$ be an integer. The matrix sector function was introduced in [19] as a generalization of the matrix sign function. The matrix sector function of $A$ can be defined as

$$\text{sect}_p(A) = A(A^p)^{-1/p}$$

where $(A^p)^{1/p}$ is the principal $p$th root (see [9]) of the matrix $A^p$. For $p = 2$ the matrix sector function is the matrix sign function [13]. In his 2008 book [9] Nicholas Higham remarked on page 49 that “a good numerical method for computing the matrix sector function is currently lacking”.

Let $k$ and $m$ be non-negative integers. A Padé approximant $[k/m]$ to the complex scalar function $f(z)$ is a rational function of the form $P_{km}(z)/Q_{km}(z)$, where $P_{km}$ and $Q_{km}$ are polynomials of degree less than or equal to $k$ and $m$, respectively, $Q_{km}(0) = 1$ and

$$f(z) - \frac{P_{km}(z)}{Q_{km}(z)} = O(z^{k+m+1})$$

as $z \to 0$.

For given $k$ and $m$ if the Padé approximant $[k/m]$ exists, then it is unique. It is usually required that $P_{km}$ and $Q_{km}$ have no common zeros, so that $P_{km}$ and $Q_{km}$ are unique (see [9, p.79]). Detailed account of the theory of Padé approximation can be found in [3].

† Corresponding author
‡ Institute of Applied Mathematics, 7 Radio Street, Vladivostok, 690041, Russia (dmkrp@yandex.ru).
∗ Department of Mathematics and Statistics, University of Regina, Regina, S4S 0A2, Canada (mлин87@ymail.com)
Let now \( P_{km}(z)/Q_{km}(z) \) be the Padé approximant \([k/m]\) of the function \((1 - z)^{-1/p}\). Since \((1 - z)^{-1/p} = 2F_1(1/p, 1; 1; z)\), where \(2F_1\) is the Gauss hypergeometric function, we may apply the formulas for the Padé approximants to this function from [2, p.65] or [20] yielding

\[
Q_{km}(z) = \sum_{j=0}^{m} \frac{(-m)_j(-\frac{1}{p} - k)_j}{j!(-k - m)_j} z^j = 2F_1(-m, -\frac{1}{p} - k; -k - m; z),
\]
\[
P_{km}(z) = \sum_{j=0}^{k} \frac{(-k)_j(\frac{1}{p} - m)_j}{j!(-k - m)_j} z^j = 2F_1(-k, -m + \frac{1}{p}; -k - m; z),
\]

(1.1)

where we assume \(k \geq m - 1 \geq 0\) are integers and \((\alpha)_k = \alpha \cdot (\alpha + 1) \cdots (\alpha + k - 1)\), \((\alpha)_0 = 1\). We will rederive this formulas in Lemma 2.1 below simultaneously finding the approximation error.

The scalar Padé iteration for the function \(\frac{\lambda}{\sqrt[2p]{x}}\), where \(\sqrt[p]{\cdot}\) denotes the principal \(p\)th root, corresponding to the Padé approximant \([k/m]\), has the form

\[
x_{l+1} = h_{km}(x_l) := x_l \frac{P_{km}(1 - x_l^p)}{Q_{km}(1 - x_l^p)}, \quad x_0 = \lambda.
\]

(1.2)

The swap between the scalar \(p\)-sector function and the \(p\)th root holds also in matrix settings and therefore from the Padé family for the matrix sector function, one obtains a family of iterations for the matrix \(p\)th root [16 formula (36)]

\[
X_{l+1} = X_l P_{km}(I - A^{-1}X_l^p)Q_{km}(I - A^{-1}X_l^p)^{-1}, \quad X_0 = I.
\]

Computation of matrix \(p\)th root has aroused considerable interest recently, see for example [5, 7, 8, 10, 11].

The following is a sample of iteration functions \(h_{km}(x)\) from the scalar Padé family (1.2):

\[
\begin{align*}
h_{00} &= x, & h_{01} &= \frac{px}{x^p + (p - 1)} \\
h_{10} &= \frac{x}{p}[(-x^p + (1 + p)], & h_{11} &= x \frac{(p - 1)x^p + (p + 1)}{(p + 1)x^p + (p - 1)} \\
h_{02} &= \frac{2p^2x}{(-p + 1)x^{2p} + (4p - 2)x^p + (2p^2 - 3p + 1)} \\
h_{12} &= \frac{x}{2p^2} \frac{[(p + 1)x^{2p} - (4p + 2)x^p + (2p^2 + 3p + 1)]}{(p + 1)x^{2p} + (4p^2 + 2p - 2)x^p + (2p^2 - 3p + 1)} \\
h_{20} &= \frac{x}{2p^2} \frac{[(p + 1)x^{2p} - (4p + 2)x^p + (2p^2 + 3p + 1)]}{(p + 1)x^{2p} - (4p + 2)x^p + (2p^2 - 3p + 1)} \\
h_{12} &= \frac{x}{2p} \frac{(-p + 1)x^{2p} + x^p(4p^2 - 2p - 2) + (2p^2 + 3p + 1)}{(2p + 1)x^p + (p - 1)} \\
h_{22} &= \frac{x}{2p} \frac{[(2p^2 - 3p + 1)x^{2p} + (8p^2 - 2)x^p + (2p^2 + 3p + 1)]}{(2p^2 + 3p + 1)x^{2p} + (8p^2 - 2)x^p + (2p^2 - 3p + 1)}.
\end{align*}
\]

The Padé approximant [1/1] provides the Halley method for the sector function that was considered in [14, 15]. B. Laszkiewicz et al [16] have proved the following relation for \(h_{01}\).
Let $x_0 = \lambda$ lie in the region $L_p^{(P_{\text{ade}})} := \{z \in \mathbb{C} : |1 - z^p| < 1\}$ and let the sequence $\{x_1\}$ be generated $h_{01}$. Then

$$|1 - x_1^p| \leq |1 - x_0^p|^2.$$ 

They also posed the following conjecture [16 Conjecture 4.2].

**Conjecture 1.2.** Let $x_i$ be the sequence generated by (1.2) for $k \geq m - 1$. If $x_0$ lies in the region $L_p^{(P_{\text{ade}})}$, then

$$|1 - x_i^p| \leq |1 - x_0^p|^{(k + m + 1)}.$$ 

(1.3)

The main purpose of this paper is to demonstrate the validity of this conjecture.

## 2. Main Results

We start with some useful lemmas.

**Lemma 2.1.** The following relation holds true:

$$\frac{P_{km}(t)}{Q_{km}(t)} = (1 - t)^{-\frac{1}{p}} - \frac{k!m!\left(\frac{1}{p}\right)_{k+1}(1 - \frac{1}{p})_{m}}{(k + m)!(k + m + 1)!} t^{k + m + 1} R_{km}(t),$$

(2.1)

where $R_{km} = 2F_1\left(m + 1, k + \frac{1}{p} + 1 \middle| \frac{1}{p} + 2m + 2 \middle| t\right)$.

**Proof.** In order to find the relation between $P_{km}(t)$ and $Q_{km}(t)$ we want to apply Euler’s transformation which reads [11 formula (2.27)]

$$2F_1(a, b; c; x) = (1 - x)^{c-a-b} 2F_1(c - a, c - b, c; x).$$

However, since $P_{km}(t), Q_{km}(t)$ are polynomials, we cannot use this formula directly. To find the right modification, choose any $\delta \in (0, 1)$ so that

$$2F_1(-k - \delta, \frac{1}{p} - m; -k - m - \delta; t) = (1 - t)^{-\frac{1}{p}} 2F_1(-m, \frac{1}{p} - k - \delta; -k - m - \delta; t).$$

Take limit $\delta \to 0$ on both sides. On the right hand side, we immediately get $(1 - t)^{-\frac{1}{p}} Q_{km}(t)$.

On the left we have

$$2F_1(-k - \delta, \frac{1}{p} - m; -k - m - \delta; t) = 1 + \frac{(-k - \delta)(\frac{1}{p} - m)}{-k - m - \delta} t + \cdots + \frac{(-k - \delta)(-k - \delta + 1) \cdots (-\delta)(\frac{1}{p} - m) - (\frac{1}{p} - m + k)}{(-k - m - \delta)(-k - m - \delta + 1) \cdots (-m - \delta)(k + 1)!} t^{k + 1} \cdots$$

$$+ \frac{(-k - \delta)(-k - \delta + 1) \cdots (-\delta + 1) \cdots (-\delta + m)(\frac{1}{p} - m) - (\frac{1}{p} - m + k)}{(-k - m - \delta)(-k - m - \delta + 1) \cdots (-m - \delta)(k + m + 1)!} t^{k + m + 1} \cdots$$

Taking limits $\delta \to 0$ we get

$$(1 - t)^{-\frac{1}{p}} Q_{km}(t) = \frac{k!m!(\frac{1}{p} - m)_{k + m + 1}}{(k + m + 1)!} t^{k + m + 1} \cdots$$

$$= P_{km}(t) + (-1)^m \frac{k!m!(\frac{1}{p} - m)_{k + m + 1}}{(k + m + 1)!} t^{k + m + 1} 2F_1\left(m + 1, k + \frac{1}{p} + 1 \middle| \frac{1}{p} + 2m + 2 \middle| t\right).$$
By using \((\frac{1}{p} - m)_{k+m+1} = (-1)^m(\frac{1}{p})_{k+1}(1 - \frac{1}{p})_m\) we obtain the identity \(\circ\). □

According to Euler’s integral representation of the Gauss hypergeometric function \([1, \text{ Theorem 2.2.1}]\) and Euler’s reflection formula \(\Gamma(z)\Gamma(1 - z) = \pi/\sin(\pi z) \ [1, \text{ Theorem 1.2.1}]\) we have

\[
(1 - t)^{-\frac{1}{p}} = 2F_1(1/p,1;1; t) = \frac{\sin(\pi/p)}{\pi} \int_0^1 u^{\frac{1}{p}-1}(1-u)^{-\frac{1}{p}} du.
\] (2.2)

**Lemma 2.2.** All the Taylor coefficients of \(\frac{P_{km}(t)}{Q_{km}(t)}\) are positive for \(k \geq m - 1 \geq 0\).

**Proof.** First we note that representation \(\circ\) implies that the function \((1 - t)^{-\frac{1}{p}}\), which is defined in the whole complex plane except a cut \([1, \infty]\) on the positive real axis, is a Stieltjes function \([3, \text{ formula (5.1.1)}]\). Since we have required \(k \geq m - 1\), then according to \([3, \text{ Theorem 5.2.1}]\) \(\frac{P_{km}(t)}{Q_{km}(t)}\) has simple positive poles lying in \((1, \infty)\) with positive residues. Summing up, we have

\[
\frac{P_{km}(t)}{Q_{km}(t)} = R(t) + \sum_{j=1}^m \frac{\lambda_j}{1 - a_j t}
\]

with \(\lambda_j > 0, 0 < a_j < 1\) and \(R(t)\) is a polynomial of degree \(k - m\) (or zero if \(k = m - 1\)). This formula makes it clear that all power series coefficients of \(\frac{P_{km}(t)}{Q_{km}(t)}\) are positive possibly except the first \(k - m + 1\) influenced by the polynomial \(R(t)\). But those are also positive as we have proved in Lemma \(\circ\). □

We next present an identity which might be of interest in its own right.

**Theorem 2.3.** The following identity holds true:

\[
(t(a + b - 1) - c + 1)2F_1\left(\begin{array}{c}
a, b \\ c \end{array} \right| t \right) 2F_1\left(\begin{array}{c}
1 - a, 1 - b \\ 2 - c \end{array} \right| t) \\
+ \frac{t(1 - t)(1 - a)(1 - b)}{2 - c} 2F_1\left(\begin{array}{c}
a, b \\ c \end{array} \right| t \right) 2F_1\left(\begin{array}{c}
2 - a, 2 - b \\ 3 - c \end{array} \right| t) \\
- \frac{t(1 - t)ab}{c} 2F_1\left(\begin{array}{c}
a + 1, b + 1 \\ c + 1 \end{array} \right| t \right) 2F_1\left(\begin{array}{c}
1 - a, 1 - b \\ 2 - c \end{array} \right| t) = 1 - c.
\] (2.3)

**Proof.** Write

\[
F(t) = 2F_1\left(\begin{array}{c}
a, b \\ c \end{array} \right| t \right), \quad G(t) = 2F_1\left(\begin{array}{c}
1 - a, 1 - b \\ 2 - c \end{array} \right| t) .
\]

Then identity \(\circ\) takes the form

\[
\Phi(t) := (t(a + b - 1) - c + 1)FG + t(1 - t)[FG' - GF'] = 1 - c.
\]

Since \(F(0) = G(0) = 1\) all we need to prove is \(\Phi'(t) = 0\). Differentiation yields:

\[
\Phi'(t) = (a + b - 1)FG + (2 - c - (3 - a - b)t)FG' - (c - (a + b + 1)t)F'G + t(1 - t)FG'' - t(1 - t)GF'' .
\]
The hypergeometric differential equation reads:
\[ t(1-t)F'' + (c - (a + b + 1)t)F' - abF = 0 \]
for \( F \) and
\[ t(1-t)G'' + (2 - c - (3 - a - b)t)G' - (1-a)(1-b)G = 0 \]
for \( G \). Simple rearrangement of the expression for \( \Phi'(t) \) then gives:
\[
\Phi'(t) = -G'[t(1-t)F'' + (c - (a + b + 1)t)F' - abF] - abFG \\
+ F'[t(1-t)G'' + (2 - c - (3 - a - b)t)G' - (1-a)(1-b)G] + (1-a)(1-b)FG + (a+b-1)FG \\
= FG(-ab + (1-a)(1-b) + (a+b-1)) = 0.
\]

\( \square \)

**Remark.** Identity (2.3) is related to several well-known formulas for hypergeometric functions such as Legendre’s identity, Elliott’s identity and Anderson-Vamanamurthy-Vuorinen’s identity. See [4] for details.

Now taking \( a = -m, b = -k - \frac{1}{p}, c = -k - m \) in the identity (2.3) we get
\[
(p(k + m + 1)(1-t) - t)Q_{km}(t)R_{km}(t) + pt(1-t)Q_{km}(t)R_{km}(t) - pt(1-t)Q'_{km}(t)R_{km}(t) = p(k + m + 1).
\]

Rewriting formula (2.1) from Lemma 2.1 in the form
\[
P_{km}(t) = (1-t)^{-\frac{1}{p}}Q_{km}(t) - \frac{k!m!(\frac{1}{p})_k(1 - \frac{1}{p})_m}{(k+m)!(k+m+1)!} t^{k+m+1} R_{km}(t) \tag{2.4}
\]
and substituting (2.4) for \( P_{km}(t) \) we obtain after some simple algebra
\[
P_{km}(t)Q_{km}(t) + p(t-1)[P'_{km}(t)Q_{km}(t) - P_{km}(t)Q'_{km}(t)]
= \frac{k!m!(\frac{1}{p})_k(1 - \frac{1}{p})_p}{(k+m)!(k+m+1)!} t^{k+m} \times \]
\[
\{ (p(k + m + 1)(1-t) - t)Q_{km}(t)R_{km}(t) + pt(1-t)Q_{km}(t)R_{km}(t) - pt(1-t)Q'_{km}(t)R_{km}(t) \}
= \frac{pk!m!(\frac{1}{p})_k(1 - \frac{1}{p})_m}{[(k+m)!]^2} t^{k+m}.
\]

This leads to the following statement.

**Theorem 2.4.** Suppose \( k \geq m - 1 \geq 0 \) and let
\[
f_{km}(t) = 1 - (1-t) \left( \frac{P_{km}(t)}{Q_{km}(t)} \right)^p := \sum_{i=0}^{\infty} c_{km,i} t^i, \tag{2.5}
\]
Then
\[
c_{km,0} = \cdots = c_{km,k+m} = 0, \ c_{km,i} > 0 \text{ for } i \geq k + m + 1.
\]
Proof. It is easy to see $f_{km}(0) = 0$, so the constant term $c_{km,0} = 0$. The desired conclusion follows from the observation that
\[
f'_{km}(t) = \left( \frac{P_{km}(t)}{Q_{km}(t)} \right)^{p-1} \cdot \frac{1}{(Q_{km}(t))^2} \cdot \left[ P_{km}(t)Q_{km}(t) + p(t-1)(P'_{km}(t)Q_{km}(t) - P_{km}(t)Q'_km(t)) \right]
\]
\[
= \left( \frac{P_{km}(t)}{Q_{km}(t)} \right)^{p-1} \cdot \frac{1}{(Q_{km}(t))^2} \cdot \frac{pk!m!(\frac{1}{p})_{k+1}(1 - \frac{1}{p})_m}{[(k + m)!]^2} \cdot t^{k+m}
\]
by the formula preceding this theorem. The expression on the right has positive Taylor coefficients starting from the term $t^{k+m}$ by Lemma 2.2. □

Remark. The case $[1/1]$ has been proved in [17] using a different approach.

Now we are in the position to give a proof to Conjecture 1.2

Proof. Let $x_0 \in L_p^{(Padé)}$ and let $t = 1 - x_0^p$. Then $|t| < 1$. From (1.2) we obtain
\[
1 - x_0^p = 1 - x_0^p \left( \frac{P_{km}(1 - x_0^p)}{Q_{km}(1 - x_0^p)} \right)^p = 1 - (1 - t) \left( \frac{P_{km}(t)}{Q_{km}(t)} \right)^p = f_{km}(t).
\]

Since $f_{km}(t)$ has non-negative Taylor coefficients by Theorem 2.3 it is clear that
\[
\max_{|t|=1} |f_{km}(t)| = f_{km}(1) = 1.
\]
Then by Schwarz’s lemma [3, Theorem 5.4.3] for all $x_0 \in L_p^{(Padé)}$
\[
|1 - x_0^p| = |f_{km}(1 - x_0^p)| < |1 - x_0^p|^{k+m+1}
\]
so that by an induction argument we see that (1.3) holds. This completes the proof. □

In the next theorem we show that the speed of convergence is in fact even higher then conjectured in [16] and proved above. Hence we have a strengthening of the Conjecture 1.2

Theorem 2.5. Let $x_l$ be the sequence generated by (1.2) for $k \geq m - 1$. If $x_0 \in L_p^{(Padé)}$, then
\[
|1 - x_l^p| \leq |1 - x_0^p|^{(k+m+1)!} \left( \frac{|1 - x_0^p + \alpha}{1 + \alpha |1 - x_0^p|} \right)^{(k+m+1)!/(k+m)}
\]
where
\[
\alpha = \frac{pk!m!(\frac{1}{p})_{k+1}(1 - \frac{1}{p})_m}{(k + m)!(k + m + 1)!} < 1.
\]
Proof. Follow the proof of Conjecture 1.2 up to application of Schwarz’s lemma. Let $\alpha$ be the first non-zero coefficient of $f_{km}$. Then apply [18, Lemma 2] yielding ($t_l = 1 - x_l^p$)

$$|t_l| = |f_{km}(t_{l-1})| < |t_{l-1}|^{k+m+1} \frac{|t_{l-1}| + \alpha}{1 + \alpha|t_{l-1}|} < \left( |t_{l-2}|^{k+m+1} \frac{|t_{l-2}| + \alpha}{1 + \alpha|t_{l-2}|} \right) \frac{|t_{l-1}| + \alpha}{1 + \alpha|t_{l-1}|} < \cdots$$

where we have used the monotonicity of $t \rightarrow (t + \alpha)/(1 + \alpha t)$ on $(0, 1)$ in the ultimate inequality. Summing up the geometric progression in the exponent in the last line gives (2.7). The value of $\alpha$ is found from (2.6). □

Remark. The value of $\alpha$ is quite small even for moderate values of $k$ and $m$. For instance, for the Halley method $k = m = 1$ we have $\alpha = (p+1)(p-1)/(12p^2)$. This shows that Theorem 2.5 is a substantial improvement over Conjecture 1.2 as well as over Theorem 1.1.

Acknowledgements. The first named author acknowledges the financial support of the Russian Basic Research Fund (grant 11-01-00038-a) and the Far Eastern Branch of the Russian Academy of Sciences (grant 09-III-A-01-008). The second named author would like to thank Dr. Fang Ren for encouragement and inspiration in his mathematical analysis course which was taken several years ago.

References

[1] G.E. Andrews, R. Askey, R. Roy, Special functions. Encyclopedia of Mathematics and its Applications, Cambridge University Press, 71 (1999).
[2] G.A. Baker Jr, Essentials of Padé Approximants. Academic Press, 1975.
[3] G.A. Baker Jr, P. Graves-Morris, Padé Approximants. Encyclopedia of Mathematics and its Applications, vol. 59. Cambridge University Press, Cambridge, 1996.
[4] R. Balasubramanian, S. Naikb, S. Ponnusamy and M. Vuorinen, Elliott’s identity and hypergeometric functions, J. Math. Anal. Appl., Vol. 271, 1(2002), 232–256.
[5] D. A. Bini, N. J. Higham, and B. Meini, Algorithms for the matrix $p$th root, Numer. Algorithms, 39 (2005) 349-378.
[6] J. Gilewicz, Story of rational approximation for the class of Stieltjes functions: from Stieltjes to recent optimal estimations of errors, Ukr. Math. J., 46 (1994) 1031–1034.
[7] C.-H. Guo, N.J. Higham, A Schur-Newton method for the matrix $p$th root and its inverse, SIAM J. Matrix Anal. Appl., 28 (2006) 788-804.
[8] C. -H. Guo, On Newton’s method and Halley’s method for the principal $p$th root of a matrix, Linear Algebra Appl., 432 (2010) 1905-1922.
[9] N.J. Higham, Functions of Matrices: Theory and Computation, SIAM, Philadelphia, 2008.
[10] B. Iannazzo, On the Newton method for the matrix $p$th root, SIAM J. Matrix Anal. Appl. 28 (2006) 503–523.
[11] B. Iannazzo, A family of rational iterations and its applications to the computation of the matrix $p$th root, SIAM J. Matrix Anal. Appl. 30 (2008) 1445–1462.
[12] C. Kenney, A.J. Laub, Padé error estimates for the logarithm of a matrix. Internat. J. Control 1989; 50(3):707–730.
[13] C. Kenney, A. J. Laub. Rational iterative methods for the matrix sign function. SIAM J. Matrix Anal. App. 12(2):273-291, 1991.
[14] C.K. Koc, B. Bakkalıoğlu, B. Halley method for the matrix sector function. IEEE Trans. Automat. Control (1995) 40(5):944–949.
[15] B. Laszkiewicz, K. Zietak, Algorithms for the matrix sector function, Electron. Trans. Numer. Anal., 31 (2008) 358–383.
[16] B. Laszkiewicz, K. Zietak, A Padé family of iterations for the matrix sector function and the matrix $p$th root, Numer. Linear Algebra Appl. 16 (2009) 951–970.
[17] M. Lin, A residual recurrence for Halley’s method for the matrix $p$th root, Linear Algebra Appl., 432 (2010) 2928-2930.
[18] R. Osserman, A sharp Schwarz inequality on the boundary, Proc. Amer.Math. Soc. vol.128, 12 (2000), 3513–3517.
[19] L.S. Shieh, Y.T. Tsay, C.T. Wang, Matrix sector functions and their applications to systems theory. Proceedings of IEE Control Theory and Applications 1984; 131(5):171–181.
[20] J. Wimp and B. Beckermann. Some explicit formulas for Padé approximants of ratios of hypergeometric functions. In World Scientific Series in Applicable Analysis vol 2, Contributions in Numerical Mathematics, 1993, 427-434.