NOTES ON RATIONAL MATRIX ITERATIONS

BRUNO IANNAZZO*

Abstract. Matrix fixed-point iterations defined by a rational function of a matrix A are considered. For these iterations it is proved that if the initial value X_0 is a polynomial of A, then the matrix convergence is reduced to scalar convergence.

It is shown that the principal Padé family of iterations for the matrix sign function and the matrix square root is a special case of a family of rational iterations due to Schröder. This characterization provides a family of iterations for the matrix pth root which preserve the structure of group of automorphisms associated with a scalar product. The first iteration in that family is the Halley method for which is proved a convergence result. Finally, it is shown that the Schur-Newton method for the matrix pth root previously applied to the inverse Newton iteration can be applied also to the direct one and to the Halley method.

Key words. matrix iteration, matrix root, matrix function, Newton's method, rational iterations, structure-preserving.

AMS subject classifications. 65F30, 15A15

1. Introduction.

The study of rational iterations, which have the form $x_{k+1} = \varphi(x_k)$, where $\varphi(z)$ is a rational function, is a topic of great interest in computation, in particular for the design and analysis of root-finding algorithms. The local convergence at a fixed point z^*, such that $z^* = \varphi(z^*)$, is related to the properties of the derivatives of φ at z^*. A study of the global convergence is very difficult: the sets of initial values for which the sequence generated by a rational iteration converges to a fixed point are bounded by the so-called Julia sets [Bea91].

The generalization to the matrix case appears in the study of matrix equations and in the computation of matrix functions [Hig]. It raises problems somehow new: it is not straightforward how to define a rational matrix iteration, there can be infinite fixed points, the lack of commutativity in finite arithmetic can have effects on the convergence, and so on.

In this paper we provide general convergence results for rational matrix iterations, then we prove some properties of specific classes of rational iterations.

General results concern the one-parameter case where the iterates are rational functions of a matrix A, say $s_k(A)$. We first prove that the uniform convergence of $s_k(z)$ on a compact set containing the eigenvalues of A implies the matrix convergence, then we show that the pointwise convergence of $s_k(\lambda)$ to attractive fixed points for each eigenvalue λ of A, implies the uniform convergence on a compact neighborhood of the spectrum of A and thus matrix convergence.

Concerning specific classes, we first consider the principal Padé family introduced in [KL91] and discussed in [Hig97, HMMT04, HMMT05, FS06]. We prove that the family can be obtained by the König root-finding method applied to the polynomial $x^2 - 1$, the method goes back to a work of Schröder in 1870 [Sch70]. Second, using the characterization given above, we extend to the König family for the polynomial $x^p - 1$ a result of Higham, Mackey, Mackey and Tisseur about the property of a part of the principal Padé family of preserving the structure of group of automorphisms associated with a scalar product. Third, we show that the Halley method, that belongs to the König family, for the computation of the matrix pth root preserves the

*Dipartimento di Fisica e Matematica, Università dell'Insubria, Via Valleggio 11, 22100 Como (bruno.iannazzo@uninsubria.it).

structure described above and we prove a result on the convergence of that method. Finally, we show that a Schur-Newton method proposed in [GH06] for the inverse Newton iteration for the computation of the matrix pth root, can be adapted to the direct Newton iteration and to the Halley method with good numerical properties.

The paper is organized as follows: in Section 2 we define the rational matrix iterations and the rational matrix iterations depending on a matrix parameter; in Section 3 we discuss the convergence of one-parameter families of rational matrix iterations; in Section 4 we show the equivalence between the principal Padé iterations and the König iterations for x^2-1; in Section 5 we generate a König family of matrix iterations preserving the structure of group of automorphism; in Section 6 and 7 we prove convergence result for the Newton and Halley method and we adapt the Schur-Newton method to them.

2. Definition. Given a rational function φ, the iteration

$$\begin{cases} z_0 \in \mathbb{C}, \\ z_{k+1} = \varphi(z_k), \quad k = 0, 1, 2, \dots \end{cases} \tag{2.1}$$

is called a rational iteration. The function φ can have some poles, so that the sequence is not necessarily well defined for each z_0. We use the notation $\varphi^{[k]}$ to address the kth iterate of the function φ, i.e., $\varphi^{[1]} = \varphi$ and $\varphi^{[k+1]} = \varphi \circ \varphi^{[k]}$. A fixed point z^* of (2.1) is such that $\varphi(z^*) = z^*$ and is said *attractive* if $|\varphi'(z^*)| < 1$. For an attractive fixed point, the *immediate basin* is the connected component of the set $\{z_0 \in \mathbb{C} : z_k \to z^*\}$ which contains z^*.

The definition of rational matrix iteration is not straightforward: consider the iteration defined by

$$\varphi(z) = \frac{az+b}{cz+d}. \tag{2.2}$$

A first matrix generalization can be obtained substituting a matrix Z to the variable z and a scalar multiple of the identity, γI, to any parameter γ, therefore,

$$\varphi(Z) = (aZ + bI)(cZ + dI)^{-1}, \tag{2.3}$$

observe that $\varphi(Z)$ is a function of Z (following the usual definition of function of a matrix [HJ90, GVL96, Hig]).

A second generalization is obtained by substituting matrices to each parameter, yielding

$$\varphi(Z) = (AZ + B)(CZ + D)^{-1}, \tag{2.4}$$

where we use the convention for which uppercase letters denote square matrices. In this case $\varphi(Z)$ in general is not a function of Z, for this reason and for the lack of commutativity in the formula of (2.4), we will use the formula of (2.3) to define a rational matrix iteration.

In the following, a *rational matrix iteration* will be an iteration

$$\begin{cases} Z_0 \in \mathbb{C}^{n\times n}, \\ Z_{k+1} = \varphi(Z_k), \quad k = 0, 1, 2, \dots \end{cases} \tag{2.5}$$

where φ is a rational function.

However, the class of rational matrix iterations is not suitable to approximate generic matrix functions, since, as we will show in Remarks 3.5 and 3.6 there hold strong conditions on the limits of such sequences.

We consider a larger class of iteration that can be studied with similar techniques as the rational matrix iterations. An iteration in this larger class can be written in the form

$$\begin{cases} Z_0 = p(A), \\ Z_{k+1} = \psi(Z_k, A), \quad k = 0, 1, 2, \dots \end{cases} \tag{2.6}$$

where $\psi = \psi(t, z)$ is a two-variable rational function and p is a polynomial. In that case, for each A, the sequence Z_k defines the same sequence of rational functions $s_k(z)$ such that $s_k(A) = \psi(s_{k-1}(A), A) = Z_k$. We call an iteration in this class a *one-parameter family of rational iterations*.

The class of one-parameter families of rational iterations contains the rational matrix iterations but also a great number of iterations used for the approximation of matrix functions.

3. Convergence of matrix iterations. Consider an iteration of the class (2.6) described above. Let Z_k be the kth iterate, therefore, $Z_k = s_k(A)$, with $s_k(z)$ rational function. In the literature the problem of convergence of that matrix iteration has been somehow neglected or addressed with more assumptions like the diagonalizability of A [Hig86, Lak98].

Here, we show that the convergence of a one-parameter family of matrix iterations is strictly related to the scalar iteration on the eigenvalues of the parameter A. In particular we first show that the uniform convergence of $s_k(z)$ in a neighborhood of the spectrum of A implies the matrix convergence. Then we show that if the sequence $s_k(\lambda)$ converges for each eigenvalue λ of A to an attractive fixed point, then the sequence $s_k(z)$ converges uniformly on a neighborhood of λ.

If $M^{-1}AM = J$, then $M^{-1}Z_kM = s_k(J)$ is block diagonal for each k, in fact if $J = J_1 \oplus \cdots \oplus J_r$, then $s_k(J) = s_k(J_1) \oplus \cdots \oplus s_k(J_r)$. Therefore, by means of the similarity M, the iteration can be uncoupled into r iterations involving only functions of the Jordan blocks. The study of the convergence is thus restricted to the case in which $A = J$ is a Jordan block of arbitrary size for the eigenvalue λ.

Moreover, in view of the formula for a function of a Jordan block (see [GVL96]),

$$f(J) = \begin{bmatrix} f(\lambda_i) & f'(\lambda_i) & \dots & \frac{f^{(k-1)}(\lambda_i)}{(k-1)!} \\ & f(\lambda_i) & \ddots & \vdots \\ & & \ddots & f'(\lambda_i) \\ \bigcirc & & & f(\lambda_i) \end{bmatrix}, \tag{3.1}$$

each of the iterates is upper triangular.

A question arises naturally: if the sequence $s_k(\lambda)$, with $s_0(\lambda) = p(\lambda)$, converges for each eigenvalue of A, what can be said about the convergence of $s_k(A)$? The following easy example shows that in general scalar convergence does not imply matrix convergence.

EXAMPLE 3.1. Consider the rational iteration $x_{k+1} = \varphi(x_k)$ where $\varphi(z) = z^2$. The sequence $r_k(1) = \varphi^{[k]}(1)$ converges to 1, but it fails to converge uniformly on any neighborhood of the point 1.

Consider the matrix iteration $X_{k+1} = X_k^2$, and the starting point $X_0 = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$; the iterates are $X_k = \begin{bmatrix} 1 & 2^k \\ 0 & 1 \end{bmatrix}$ and the sequence fails to converge to the identity matrix. For this iteration and X_0 being a Jordan block of size n for the eigenvalue 1, there is matrix convergence only for $n = 1$, that is, the scalar case.

A sufficient condition for the convergence of the matrix sequence is stated in the following result, in which we use the notation $\|f(z)\|_K = \sup_{x\in K} |f(x)|$ for a compact set K.

LEMMA 3.2. *If $s_k(z)$ is a sequence of rational functions that converges uniformly in a compact neighborhood of λ to the function $f(z)$, then $s_k(J)$ converges to $f(J)$, where $J = J(\lambda, n)$ is a Jordan block of arbitrary size n relative to the eigenvalue λ. Moreover, if the convergence of $s_k(\lambda)$ to $f(\lambda)$ is of order q, then the convergence of $s_k(J)$ to $f(J)$ is dominated by a sequence converging with order q.*

Proof. Observe that from the formula (3.1), the matrix sequence converges if the sequence $s_k(z)$ and its derivatives up to the order $n-1$ converge. Consider a small circle γ of center λ and radius R, fully contained in a compact neighborhood of λ in which uniform convergence occurs, i.e., $\|s_k(z) - f(z)\|_K \to 0$. From the Cauchy formula it follows that

$$\left| \frac{s_k^{(p)}(\lambda)}{p!} - \frac{f^{(p)}(\lambda)}{p!} \right| = \left| \frac{1}{2\pi i} \oint_\gamma \frac{s_k(z) - f(z)}{(z-\lambda)^{p+1}} \right| \leqslant \frac{1}{R^p} \|s_k(z) - f(z)\|_K \to 0,$$

for $p = 0, \dots, n-1$. The previous relation provides the convergence of the sequence $s_k(J)$ to $f(J)$. It provides also the proof of the assertion about the order of convergence. In fact, from formula (3.1) it follows that

$$\|s_k(J) - f(J)\|_\infty = \sum_{p=0}^{n-1} \left| \frac{s_k^{(p)}(\lambda)}{p!} - \frac{f^{(p)}(\lambda)}{p!} \right| \leqslant \|s_k(\lambda) - f(\lambda)\|_K \sum_{p=0}^{n-1} \frac{1}{R^p}. \quad (3.2)$$

□

In conclusion, if the sequence $s_k(z)$ converges uniformly on a compact neighborhood of the spectrum of A, then the sequence $s_k(A)$ converges. If the scalar convergence is not uniform, then the matrix iteration may fail to converge, as Example 3.1 shows.

We have turned the problem from matrix convergence to uniform convergence on a compact neighborhood of the spectrum. This does not seem at first sight an advantage, but it is possible to show that uniform convergence of the sequence $s_k(z)$ occurs in most cases. Roughly speaking, a sufficient condition for the uniform convergence is that the fixed points to which the scalar iteration converges are attractive.

THEOREM 3.3. *Given a matrix iteration $X_{k+1} = \psi(X_k, A)$, and let $X_0 = p(A)$, where $\psi(t, z)$ is a rational function with respect to the two variables and $p(z)$ is a polynomial. If for each λ, eigenvalue of A, the scalar sequence*

$$\begin{cases} s_0(\lambda) = p(\lambda) \\ s_k(\lambda) = \psi(s_{k-1}(\lambda), \lambda) \end{cases}$$

converges to a complex number $f(\lambda)$ and is such that $\left| \dfrac{\partial \psi}{\partial t}(t, z)|_{(f(\lambda),\lambda)} \right| < 1$, then the matrix iteration converges.

Proof. In view of Lemma 3.2, it is enough to prove that the scalar iteration, starting with $s_0 = p(z)$, gives a sequence which converges uniformly to a function $f(z)$ for each z in a compact neighborhood of the spectrum of A.

The proof is based on three steps: first, it is proved that the iteration is a contraction for each z in a compact neighborhood of an eigenvalue λ of A, this enables us to define the function of the fixed points $f(z)$; then, it is proved that this function is continuous at λ; finally, it is proved that the convergence to this function is uniform in a compact neighborhood of λ.

1) The function $g(t,z) \doteq \frac{\partial \psi}{\partial t}(t,z)$ is continuous in a neighborhood of $(f(\lambda),\lambda)$, thus, there exist R_1 and ρ_1 such that for $|t - f(\lambda)| \leqslant R_1$ and $|z-\lambda| \leqslant \rho_1$, one has $|g(t,z)| \leqslant M < 1$.

For each z such that $|z-\lambda| \leqslant \rho_1$, the function $t \to g(t,z)$ is a contraction on the set $|t - f(\lambda)| \leqslant R_1$, where it has a fixed point that depends on z, say $f(z)$.

2) Observe that, for each z such that $|z-\lambda| \leqslant \rho_1$, there exist $\vartheta_1, \vartheta_2 \in (0,1)$, such that for $\xi = f(z) + \vartheta_1(f(\lambda) - f(z))$, $\eta = z + \vartheta_2(\lambda - z)$, it holds that

$$\begin{aligned} |f(z) - f(\lambda)| &= |\psi(f(z),z) - \psi(f(\lambda),\lambda)| \\ &= \left| \frac{\partial \psi}{\partial t}(t,z)\big|_{(\xi,\eta)} (f(z) - f(\lambda)) \right| + \left| \frac{\partial \psi}{\partial z}(t,z)\big|_{(\xi,\eta)} (z-\lambda) \right| \\ &\leqslant M|f(z) - f(\lambda)| + N|z-\lambda|, \end{aligned}$$

where N is the maximum of $\left|\frac{\partial \psi}{\partial z}(t,x)\right|$ over $|t - f(\lambda)| \leqslant R_1$ and $|z-\lambda| \leqslant \rho_1$.

Since $M < 1$, the inequality

$$|f(z) - f(\lambda)| \leqslant \frac{N}{1-M}|z-\lambda|$$

proves the continuity of f at λ.

3) Since $s_k(\lambda) \to f(\lambda)$, there exists $\bar{k}$ such that $|s_{\bar{k}}(\lambda) - f(\lambda)| \leqslant R_1/3$ and for the continuity of $s_{\bar{k}}$ and f there exists ρ_2 such that for $|z-\lambda| \leqslant \rho_2$, $|s_{\bar{k}}(z) - s_{\bar{k}}(\lambda)| \leqslant R_1/3$ and $|f(z) - f(\lambda)| \leqslant R_1/3$.

Thus, for $|z-\lambda| \leqslant \rho_2$, one has

$$|s_{\bar{k}}(z) - f(z)| \leqslant |s_{\bar{k}}(z) - s_{\bar{k}}(\lambda)| + |s_{\bar{k}}(\lambda) - f(\lambda)| + |f(\lambda) - f(z)| \leqslant R_1.$$

This formula can be used to prove, by an induction argument, that for each integer h, $|s_{\bar{k}+h}(z) - f(z)| \leqslant R_1$ and

$$|s_{\bar{k}+h}(z) - f(z)| \leqslant M^h R_1 \to 0.$$

Taking the maximum over the set $|z-\lambda| \leqslant \rho_2$, one has the uniform convergence of $s_k(z)$ to $f(z)$. □

In the case $\psi(z,t) = \varphi(z)$, the convergence of the above theorem is to a locally constant function, this leads to the following.

COROLLARY 3.4. *Let $X_{k+1} = \varphi(X_k)$ be a rational matrix iteration. If, for each eigenvalue λ of X_0, the scalar iteration $x_{k+1} = \varphi(x_k)$ with $x_0 = \lambda$, converges to an attractive fixed point, then $\lim_k X_k$ is diagonalizable.*

Proof. Define $s_k(z) = \varphi^{[k]}(z)$ and let λ be an eigenvalue of X_0 such that $s_k(\lambda)$ converges to an attractive fixed point z^*. The sequence $s_k(z)$ converges to the constant function $f(z) = z^*$ uniformly on a compact neighborhood of z. This implies matrix convergence to $f(J) = z^* I$, for each Jordan block J relative to the eigenvalue λ.

Repeating this argument for each eigenvalue of X_0 and using the Jordan decomposition, the proof is achieved. □

REMARK 3.5. Corollary 3.4 can be used in two ways: first, to prove that a matrix function defined as the limit of a rational iteration is diagonalizable, for instance, it is an alternative way to prove the well-known fact that the matrix sign function is diagonalizable; second, to prove that there does not exist a rational iteration that converges to a function that is not diagonalizable, for instance, there does not exist a rational iteration converging to the matrix pth root, logarithm or exponential.

REMARK 3.6. The fixed points of a matrix iteration are the solutions of the equation $\varphi(X) = X$. If $\varphi(z) = p(z)/q(z)$, one has that the fixed points are the matrices X such that $Xq(X) - p(X) = 0$. The solutions of this equation have eigenvalues among the zeros of the polynomial $zq(z) - p(z)$, which are at most d where $d-1$ is the degree of the rational iteration. From that, it follows that a lower bound for the minimal degree of an iteration that converges to a matrix function is the number of eigenvalues of the matrix function. This is another proof that matrix functions like $A^{1/p}$ or $\exp(A)$ cannot be approximated by rational iterations, since they can have an arbitrary number of distinct eigenvalues.

4. Equivalence between the König family and the principal Padé iterations family. In the paper [KL91] Kenney and Laub derive a family of rational iterations for the computation of the matrix sign function. The derivation is based on the theory of Padé approximations and exploits the relation

$$\operatorname{sign}(z) = \frac{z}{\sqrt{z^2}} = \frac{z}{\sqrt{1-(1-z^2)}} = \frac{z}{\sqrt{1-\xi}},$$

where $\xi = 1 - x^2$.

They consider the approximants of the function

$$h(\xi) = (1-\xi)^{-1/2},$$

which are well known because h is a particular case of a hypergeometric function.

Given $p_{mn}(\xi)/q_{mn}(\xi)$, the (m,n) Padé approximant to h, Kenney and Laub set

$$x_{k+1} = f_{mn}(x_k) = x_k \frac{p_{mn}(1-x_k^2)}{q_{mn}(1-x_k^2)}$$

and obtain a family of iterations for the matrix sign function.

The most interesting iterations for their good convergence properties are the ones relative to the cases $m = n-1$ and $m = n$ which have been called *principal Padé iterations* [Hig]. For these values of m and n one can define

$$g_r(x) = f_{mn}(x), \text{ for } r = m+n+1, \tag{4.1}$$

for which the following result holds [KL91].

THEOREM 4.1. *For the function* (4.1) *it holds that:*

1. *the iteration $x_{k+1} = g_r(x_k)$ is convergent for each nonimaginary x_0, with order of convergence r;*
2. $g_r(x) = \dfrac{(1+x)^r - (1-x)^r}{(1+x)^r + (1-x)^r}$.

Higham [Hig] noticed that these families were essentially derived by Howland [How83], though for even r the iteration functions of Howland are the reciprocal of the ones of Kenney and Laub.

We claim that the family of principal Padé iterations goes back to Schröder in its monumental paper of 1870 [Sch70] (an English translation is available in [Sch93]). This family was studied by Householder [Hou53] and many other authors, who called it König family or basic family.

The König's method of order σ, applied to the function f, is defined by the formula [BH03]

$$K_{f,\sigma}(z) = z + (\sigma - 1)\frac{(1/f(z))^{(\sigma-2)}}{(1/f(z))^{(\sigma-1)}}, \tag{4.2}$$

where $(1/f)^{(k)}$ is the kth derivative of $1/f$. It can be proved that the method converges to simple roots of f with order at least σ. For $\sigma = 2$ the König method is the Newton method, for $\sigma = 3$ is the so-called Halley method.

If f is a polynomial, then $K_{f,\sigma}$ is a rational function. Let us define $K_{p,\sigma}$ as the König family applied to the polynomial $f = x^p - 1$.

THEOREM 4.2. *For the König rational functions relative to the polynomial $x^2 - 1$ it holds that $K_{2,r}(x) = \dfrac{(x+1)^r + (x-1)^r}{(x+1)^r - (x-1)^r}$. Thus, $K_{2,r}$ coincides with g_r of (4.1) for odd r and with the reciprocal of g_r for even r.*

Proof. From

$$\frac{d^n}{dx^n}\left(\frac{1}{x^2-1}\right) = \frac{1}{2}\frac{d^n}{dx^n}\left(\frac{1}{x-1} - \frac{1}{x+1}\right)$$
$$= \frac{(-1)^n n!}{2}\left(\frac{1}{(x-1)^{n+1}} - \frac{1}{(x+1)^{n+1}}\right) = \frac{(-1)^n n!}{2}\left(\frac{(x+1)^{n+1} - (x-1)^{n+1}}{(x^2-1)^{n+1}}\right),$$

it follows that

$$K_{2,r}(x) = x - (x^2-1)\frac{(x+1)^{r-1} - (x-1)^{r-1}}{(x+1)^r - (x-1)^r} = \frac{(x+1)^r + (x-1)^r}{(x+1)^r - (x-1)^r}.$$

□

5. Structure-preserving algorithms in the König family. In [HMMT05] it has been proved that an iteration preserving the structure of group of automorphisms of a scalar product is of the form

$$\pm z^k \frac{q(z)}{\operatorname{rev} q(z)}, \tag{5.1}$$

where $\operatorname{rev} q(z) = z^d q(1/z)$, for a given polynomial $q(z)$ of degree d. To ease the notation we say that an iteration is *structure-preserving* if it preserves the structure of group of automorphism of a scalar product.

The principal Padé iterations and, in view of Theorem 4.2, the $K_{2,\sigma}$ iterations, for odd σ, are iterations for the matrix sign function which are structure-preserving [HMMT05]; this is a case of a more general theorem.

THEOREM 5.1. *If $n \equiv 3 \pmod{p}$, then the function $K_{p,n}$ has the form*

$$z\frac{q(z)}{\operatorname{rev} q(z)}, \tag{5.2}$$

i.e., is structure-preserving.

Proof. The proof is obtained by computing a formula for the derivative of $1/(x^p-1)$ and, from it, an explicit elementary formula for the König function from which we deduce the theorem.

Let $\omega = \sin(2\pi/p) + \mathbf{i}\cos(2\pi/p)$ and $\varphi(x) = (x^p-1)/(x-1) = \sum x^k$. Observe that $\varphi(\omega^k) = 0$ for $k \not\equiv 0 \pmod p$.

It holds that

$$\frac{1}{x^p-1} = \frac{1}{p}\sum_{k=0}^{p-1}\frac{\omega^k}{x-\omega^k},$$

in fact,

$$\sum_{k=0}^{p-1}\frac{\omega^k}{x-\omega^k} = \frac{1}{x^p-1}\sum_{k=0}^{p-1}\frac{1}{\bar\omega^k}\frac{x^p-1}{x-\omega^k} = \frac{1}{x^p-1}\sum_{k=0}^{p-1}\frac{(\bar\omega^k x)^p-1}{\bar\omega^k x-1}$$
$$= \frac{1}{x^p-1}\sum_{k=0}^{p-1}\varphi(\bar\omega^k x) = \frac{1}{x^p-1}\sum_{k=0}^{p-1}\sum_{r=0}^{p-1}(\bar\omega^k x)^r = \frac{1}{x^p-1}\sum_{r=0}^{p-1}x^r\sum_{k=0}^{p-1}\bar\omega^{kr} = \frac{p}{x^p-1}.$$

Now,

$$\frac{d^n}{dx^n}\left(\frac{1}{x^p-1}\right) = \frac{1}{p}\sum_{k=0}^{p-1}\frac{d^n}{dx^n}\frac{\omega^k}{x-\omega^k} = \frac{(-1)^n n!}{p}\sum_{k=0}^{p-1}\frac{\omega^k}{(x-\omega^k)^{n+1}}$$
$$= \frac{(-1)^n n!}{(x^p-1)^{n+1}}\sum_{k=0}^{p-1}\bar\omega^{kn}\varphi^{n+1}(\bar\omega^k x) = \frac{(-1)^n n!}{(x^p-1)^{n+1}}\sum_{k=0}^{p-1}\omega^{kn}\varphi^{n+1}(\omega^k x),$$

and, defining $\psi_n(x) = \dfrac{1}{p}\displaystyle\sum_{k=0}^{p-1}\omega^{k(n-1)}\varphi^n(\omega^k x)$, it holds the explicit formula

$$K_{p,n} = x - (x^p-1)\frac{\psi_{n-1}}{\psi_n} = \frac{x\psi_n - (x^p-1)\psi_{n-1}}{\psi_n}.$$

Observe that the denominator of $K_{p,n}$, namely $\psi_n(x)$, is formed by the terms of $\varphi^n(x)$ in which the exponent of x is congruent to $(1-n)$ modulo p, in fact, if $\varphi^n(x) = \sum a_r x^r$, then

$$\psi_n(x) = \frac{1}{p}\sum_{k=0}^{p-1}\omega^{k(n-1)}\sum_r a_r\omega^{kr}x^r = \frac{1}{p}\sum_r\left(a_r x^r\sum_{k=0}^{p-1}\omega^{k(n+r-1)}\right) = \sum_{r\equiv 1-n} a_r x^r.$$

Observe that the numerator of $K_{p,n}$, namely $x\psi_n(x) - (x^p-1)\psi_{n-1}(x)$, is formed by the terms of $\varphi^n(x)$ in which the exponent of x is congruent to $(2-n)$ modulo p, in fact

$$x\psi_n(x) - (x^p-1)\psi_{n-1}(x) = \frac{1}{p}\left(\sum_{k=0}^{p-1}\omega^{k(n-1)}x\varphi^n(\omega^k x) - \omega^{k(n-2)}(x^p-1)\varphi^{n-1}(\omega^k x)\right)$$
$$= \frac{1}{p}\sum_{k=0}^{p-1}\left(\omega^{k(n-1)}x\varphi^n(\omega^k x) - \omega^{k(n-2)}(\omega^k x-1)\varphi^n(\omega^k x)\right) = \frac{1}{p}\sum_{k=0}^{p-1}\omega^{k(n-2)}\varphi^n(\omega^k x)$$
$$= \frac{1}{p}\sum_{k=0}^{p-1}\omega^{k(n-2)}\sum_r a_r\omega^{kr}x^r = \frac{1}{p}\sum_r\left(a_r x^r\sum_{k=0}^{p-1}\omega^{k(n+r-2)}\right) = \sum_{r\equiv 2-n} a_r x^r,$$

where we have used the identity $x^p - 1 = (\omega^k x - 1)\varphi(\omega^k x)$, for any k.

To conclude, let $a_{\alpha_1}, \ldots, a_{\alpha_\nu}$ be the coefficients of $\varphi^n(x)$ relative to exponents congruent to $1-n$ modulo p, and let $a_{\beta_1}, \ldots, a_{\beta_\mu}$ be the coefficients of $\varphi^n(x)$ relative to exponents congruent to $2-n$ modulo p, so that

$$K_{p,n} = \frac{a_{\alpha_1} x^{\alpha_1} + \cdots + a_{\alpha_\nu} x^{\alpha_\nu}}{a_{\beta_1} x^{\beta_1} + \cdots + a_{\beta_\mu} x^{\beta_\mu}}.$$

To conclude the proof, it is enough to prove that, for $n \equiv 3 \pmod p$, it holds that $a_{\alpha_1} = a_{\beta_\mu}, a_{\alpha_2} = a_{\beta_{\mu-1}}, \ldots, a_{\alpha_\nu} = a_{\beta_1}$. Let $N = \deg \varphi^n(x)$, observe that since $N = np - n$, then $\beta_\mu = N + 2 - p$.

Since $\varphi^n(x) = \operatorname{rev} \varphi^n(x)$, namely $a_r = a_{N-r}$ for each $r = 0, 1, \ldots$, then $a_{\alpha_1} = a_{\beta_\mu} = a_{N+2-p} = a_{p-2}$ when $p - 2 \equiv \alpha_1 \equiv 1 - n \pmod p$, i.e., $n \equiv 3 \pmod p$.

The equalities $a_{\alpha_{i+1}} = a_{\beta_{\mu-i}}$ for $i = 1, 2, \ldots$, follow from the fact that if $n \equiv 3 \pmod p$, then $\alpha_1 = p-2$ and thus $\alpha_{i+1} = (i+1)p-2 = N-(N+2-p-ip) = N-\beta_{\mu-i}$. □

By the properties of the König method [BH03], the iteration $z_{k+1} = K_{p,n}(z_k)$ converges locally, with order of convergence at least n, to the roots of the polynomial $x^p - 1$. It is easy to see, by an induction argument, that the iteration

$$x_{k+1} = a^{1/p} K_{p,n}(a^{-1/p} x_k), \tag{5.3}$$

for $x_0 = a^{1/p} z_0$, is such that $x_k = a^{1/p} z_k$ and thus converges locally to the roots of $x^p - a = 0$.

Iteration (5.3) does not seem effective for computing the pth roots of a, since it uses $a^{1/p}$, but from the proof of Theorem 5.1 it follows that there exists a real polynomial $\widehat{q}(z)$, such that

$$K_{p,n}(z) = z \frac{\widehat{q}(z^p)}{\operatorname{rev} \widehat{q}(z^p)}, \tag{5.4}$$

for $n \equiv 3 \pmod p$. For these values of n, the iteration for x_k has the form

$$x_{k+1} = x_k \frac{\widehat{q}(a^{-1} x_k^p)}{\operatorname{rev} \widehat{q}(a^{-1} x_k^p)}, \tag{5.5}$$

which is an effective iteration to approximate with a high precision the pth roots of a given number.

In matrix terms, iteration (5.5) is structure-preserving for the matrix pth root (using Theorem 3.3). In fact if a is a matrix belonging to an automorphism group and $x_0 = a$, then each x_k belongs to that group.

A difficulty in the use of iteration (5.5) is the global convergence. We will not investigate further on global convergence of (5.5), but in Section 6 we will give a convergence proof for the case $n = 3$ which is structure-preserving iteration for each p, in view of Theorem 5.1.

REMARK 5.2. We consider an application of Theorem 5.1 to the theory of root-finding algorithms. A rational iterative root-finding algorithm is said *generally convergent* if it converges to a root for almost every initial guess and for almost every polynomial (where the Lebesgue measure on the complex plane and on the space of coefficients is considered).

It is known that the Newton method is generally convergent for quadratic polynomials, but not for cubic. In fact, the Newton iteration for the polynomial $p(z) =$

z^3-2z+2 does not converge to any root for initial values in a suitable set of measure greater than zero.

McMullen has constructed in [McM87] a generally convergent algorithm for cubic polynomials and has proved that there does not exist a generally convergent algorithm for polynomial of degree greater than three.

Using the results of McMullen, Hawkins has proved that any generally convergent root-finding algorithm is generated by a root-finding algorithm for the polynomial x^3-1 of the form (5.4) [Haw02]. Thus, Theorem 5.1 can be used to construct generally convergent algorithms for cubic polynomial of arbitrarily high order of convergence.

6. Nice properties of the Halley method. The König method of order 3 is the so-called Halley method which, for the equation $x^p - a = 0$, is

$$x_{k+1} = x_k \frac{(p-1)x_k^p + (p+1)a}{(p+1)x_k^p + (p-1)a}, \quad x_0 \in \mathbb{C}. \tag{6.1}$$

Here we considered the Halley method for the computation of the principal pth root of a.

Given an iteration for the computation of the principal pth root of a number, we call *principal basin for the initial value* x_0 (or simply *principal basin* if there is no ambiguity on the initial value) the set $\{a \in \mathbb{C} : x_k \to a^{1/p}\}$.

A very nice feature of the Halley method is that if one chooses as initial value $x_0 = 1$, the *principal basin* is somewhat *larger* than the one of the Newton method: it has been proved [Ian06] that for the Newton method the principal basin for $x_0 = 1$ contains the set

$$\mathcal{D} = \{z \in \mathbb{C} : \operatorname{Re} z > 0, |z| \leqslant 1\}, \tag{6.2}$$

whence for the Halley method it holds the following result.

THEOREM 6.1. *The principal basin of iteration* (6.1) *for the initial value* $x_0 = 1$ *contains the set* $\mathbb{C}_> = \{z \in \mathbb{C} : \operatorname{Re} z > 0\}$.

With a strategy similar to the one used in [BHM05, Ian06] we can *swap* the iteration with another one more simple to work with.

PROPOSITION 6.2. *Let* $a \neq 0$ *be a complex number, then iteration* (6.1) *with the initial value* $x_0 = 1$ *converges to* $a^{1/p}$ *if and only if the iteration*

$$\begin{cases} z_{k+1} = z_k \dfrac{(p-1)z_k^p + (p+1)}{(p+1)z_k^p + (p-1)} \\ z_0 = a^{-1/p} \end{cases} \tag{6.3}$$

converges to 1.

Proof. It is sufficient to prove by induction that $z_k = x_k a^{-1/p}$. □

Theorem 6.1, in view of Proposition 6.2, can be restated in an equivalent way for iteration (6.3).

THEOREM 6.3. *The immediate basin of attraction for the fixed point* 1 *of the rational iteration* (6.3) *contains the sector*

$$\mathcal{S}_{2p} = \{z \in \mathbb{C} \setminus \{0\} : -\pi/(2p) < \arg(z) < \pi/(2p)\}.$$

Proof. Let us define

$$\varphi(z) = \frac{(p-1)z^p + (p+1)}{(p+1)z^p + (p-1)},$$

iteration (6.3) can be written as $z_{k+1} = z_k\varphi(z_k)$. The sector $\mathcal{S}_{2p}$ contains the point $z = 1$, is connected and, by Lemma 6.4, if $z \in \mathcal{S}_{2p}$ then $z\varphi(z) \in \mathcal{S}_{2p}$. Thus, the set $\mathcal{S}_{2p}$ belongs to the immediate basin of the fixed point $z = 1$. In fact, given a rational iteration $x_{k+1} = \psi(x_k)$, any connected set $\mathcal{U}$ containing only a fixed point z^* and such that $\psi(\mathcal{U}) \subset \mathcal{U}$ belongs to the immediate basin of z^* [Bea91]. □

LEMMA 6.4. *For each $z \in \mathcal{S}_{2p}$, it holds that $|\arg(z\varphi(z))| \leqslant |\arg(z)|$ and the equality holds if and only if z is real.*

Proof. If z is real then $\varphi(z)$ is real. Let us consider the case $\arg(z) > 0$, since $\arg(z\varphi(z)) = \arg(z) + \arg(\varphi(z))$, it is enough to prove that

$$-2\arg(z) < \arg(\varphi(z)) < 0. \tag{6.4}$$

Removing real positive constants, it holds that

$$\arg(\varphi(z)) = \arg\big(\,((p-1)z^p + (p+1)) \cdot ((p+1)\bar{z}^p + (p-1))\,\big).$$

Using the decomposition $z = r(\cos\vartheta + \mathbf{i}\sin\vartheta)$, one has

$$\arg(\varphi(z)) = \arg\big((p^2-1)(|r|^{2p}+1) + 2(p^2+1)r^p\cos(p\vartheta) - 4\mathbf{i}pr^p\sin(p\vartheta)\big).$$

Applying the tangent trigonometric function to the inequalities (6.4) it is obtained the equivalent

$$-\frac{\sin(2\vartheta)}{\cos(2\vartheta)} < \frac{-4pr^p\sin(p\vartheta)}{(p^2-1)(r^{2p}+1) + 2(p^2+1)r^p\cos(p\vartheta)} < 0. \tag{6.5}$$

The latter inequality is evident from $0 < \vartheta < \pi/(2p)$. The former need a bit more work and can be rewritten as

$$(p^2-1)\sin(2\vartheta)r^{2p} + 2\big((p^2+1)\cos(p\vartheta)\sin(2\vartheta) - 2p\sin(p\vartheta)\cos(2\vartheta)\big)r^p + (p^2-1)\sin(2\vartheta) > 0, \tag{6.6}$$

and can be seen as a quadratic inequality on the variable $x = r^p$. The quadratic has the form $\gamma(x) = ax^2 + 2bx + a$, where $a = (p^2-1)\sin(2\vartheta)$ and $b = (p^2+1)\cos(p\vartheta)\sin(2\vartheta) - 2p\sin(p\vartheta)\cos(2\vartheta)$. Since $a > 0$, the inequality $\gamma(x) > 0$ is true if the equation $\gamma(x) = 0$ has no solution. Observe that if $\gamma(x) = 0$ then $\gamma(1/x) = 0$ and then if $\gamma(1) > 0$ there exist no positive solution.

Using the inequalities $(\vartheta - \vartheta^3/6) \leqslant \sin\vartheta \leqslant \vartheta$ for $0 < \vartheta < \pi/(2p)$ and $\sin((p-2)\vartheta) = \sin(p\vartheta)\cos(2\vartheta) + \sin(2\vartheta)\cos(p\vartheta)$, one can see that

$$\begin{aligned}\gamma(1) \geqslant (p^2-1)\sin(2\vartheta) - 2p\sin((p-2)\vartheta) &> (p^2-1)(2\vartheta - \frac{4}{3}\vartheta^3) - 2p(p-2)\vartheta \\ &= \frac{2}{3}\vartheta\left(6p - 3 - 2(p^2-1)\vartheta^2\right),\end{aligned}$$

the last expression is positive if $\vartheta^2 \leqslant \frac{6p-3}{2p^2-2}$ and this is true since $\vartheta \leqslant \pi/(2p)$. □

Theorems 3.3 and 6.1 guarantee the convergence to $A^{1/p}$ of the matrix iteration

$$X_{k+1} = X_k\frac{(p-1)X_k^p + (p+1)A}{(p+1)X_k^p + (p-1)A}, \tag{6.7}$$

with $X_0 = I$, for each A having eigenvalues in $\mathbb{C}_>$. Moreover, for what we have proved in Section 5, iteration (6.7) is structure-preserving.

Iteration (6.7) cannot be used directly to approximate the principal pth root, in fact, using the same idea as in [Ian06], one can prove that iteration (6.7) is *unstable in a neighborhood of* $A^{1/p}$, i.e., a perturbation on the value of X_k is amplified in the following steps preventing the convergence in a finite arithmetic computation.

This problem can be overrided using another algorithm which provides the same sequence but which is stable in a neighborhood of $A^{1/p}$, for instance

$$\begin{cases} X_0 = I, \quad N_0 = A, \\ X_{k+1} = X_k((p+1)I + (p-1)N_k)^{-1}((p-1)I + (p+1)N_k), \\ N_{k+1} = N_k \left((p+1)I + (p-1)N_k)^{-1}((p-1)I + (p+1)N_k)\right)^{-p}. \end{cases} \tag{6.8}$$

where $N_k \to I$ and $X_k \to A^{1/p}$.

7. The Schur-Newton method. A one-parameter family of iterations for computing the principal pth root of a matrix A is

$$X_{k+1} = \frac{(p-1)X_k + AX_k^{1-p}}{p}, \tag{7.1}$$

which coincides with the Newton method for the equation $X^p - A = 0$, when the latter is well defined and X_0 commutes with A [Smi03].

In [Smi03] it is proved that this iteration is not stable in a neighborhood of $A^{1/p}$, the stable variant

$$\begin{cases} Y_0 = I, \quad N_0 = A, \\ Y_{k+1} = Y_k \left(\dfrac{(p-1)I + N_k}{p}\right), \\ N_{k+1} = \left(\dfrac{(p-1)I + N_k}{p}\right)^{-p} N_k, \end{cases} \tag{7.2}$$

has been proposed in [Ian06], where it has been proved that (Y_k, N_k) converges quadratically to $(A^{1/p}, I)$ for each A having eigenvalues in the set $\mathcal{D}$ of (6.2). An algorithm for computing the principal pth root is the following.

ALGORITHM 1 (A Newton method for $A^{1/p}$ [Ian06]). Given $A \in \mathbb{C}^{n\times n}$ with no nonpositive real eigenvalues, an integer $p > 2$ and an algorithm for computing the square root.

1. Compute B, the principal square root of A;
2. Set $C = B/\|B\|$ for a suitable norm. The eigenvalues of C belongs to the set $\mathcal{D}$ of (6.2);
3. By means of iteration (7.2)
   - If p is even, compute $S = C^{2/p}$, the $(p/2)$th root of C and set $X = S\|B\|^{2/p}$;
   - If p is odd, compute $S = C^{1/p}$, the pth root of C and set $X = \left(S\|B\|^{1/p}\right)^2$.

Algorithm 1 has a complexity of $O(n^3 \log p)$ arithmetics operations per step. The initial square root can be obtained by forming the Schur decomposition of A, without affecting the complexity order with respect to p. There is an important consideration to do: the Schur decomposition gives information on the eigenvalues of A and that information is not exploited in the algorithm.

Since the number of steps to achieve the required accuracy in the numerical computation depends on the localization of the eigenvalues of the matrix whose pth root is required, a smarter preprocessing would give a better conditioning and reduce the number of steps needed for the expensive iteration (7.2) (or other similar).

In order to give a better localization of the eigenvalues, one could perform a small number of initial square roots without affecting the order of complexity of the overall algorithm. Moreover, multiplying the preprocessed matrix by a scalar parameter could reduce the number of steps needed for convergence.

On these ideas is based a Schur-Newton algorithm of Guo and Higham [GH06], who consider a Newton iteration for the inverse pth root. The iteration, introduced in [BHM05], is

$$X_{k+1} = \frac{1}{p}\left((p+1)X_k - X_k^{p+1}A\right), \quad X_0 = I, \tag{7.3}$$

which converges to $A^{-1/p}$, and for which in [GH06] it is constructed a nice convergence region for the eigenvalues of A: if the spectrum of A belongs to that region, then $X_k \to A^{1/p}$. From iteration (7.3) it can be obtained a stable iteration [Ian06, Lak98]

$$\begin{cases} Y_0 = \dfrac{1}{c}I, \quad N_0 = \dfrac{1}{c^p}A, \\ Y_{k+1} = Y_k\left(\dfrac{(p-1)I - N_k}{p}\right), \\ N_{k+1} = \left(\dfrac{(p-1)I - N_k}{p}\right)^p N_k, \end{cases} \tag{7.4}$$

such that $Y_k \to A^{-1/p}$ and $N_k \to I$. Setting $X_k = Y_k^{-1}$ it is obtained the iteration [GH06]

$$\begin{cases} X_0 = cI, \quad N_0 = \dfrac{1}{c^p}A, \\ Y_{k+1} = \left(\dfrac{(p-1)I - N_k}{p}\right)^{-1} Y_k, \\ N_{k+1} = \left(\dfrac{(p-1)I - N_k}{p}\right)^p N_k, \end{cases} \tag{7.5}$$

for which $X_k \to A^{1/p}$.

Algorithm 2 (Schur-Newton algorithm using (7.4) and (7.5) [GH06]). Given $A \in \mathbb{C}^{n\times n}$ with no nonpositive real eigenvalues, an integer $p = 2^{k_0}q$ with $k_0 \geqslant 0$ and q odd.

1. Compute the Schur decomposition of $A = QRQ^T$;
2. If $q = 1$ then $k_1 = k_0$ else choose $k_1 \geqslant k_0$ such that $\arg(\lambda_1^{1/2^{k_1}}) \in (-\pi/8, \pi/8)$ and $|\lambda_1/\lambda_n|^{1/2^{k_1}} \leqslant 2$, where the eigenvalues of A are ordered $|\lambda_n| \leqslant \cdots \leqslant |\lambda_1|$;
3. Compute $B = R^{1/2^{k_1}}$ by squaring k_1 times; if $q = 1$, then $X = B$, else continue;
4. Let $\mu_1 = |\lambda_1|^{1/2^{k_1}}$, $\mu_n = |\lambda_n|^{1/2^{k_1}}$;
   - If the λ_i are all real, if $\mu_1 \neq \mu_n$ determine $c = \left(\frac{\alpha^{1/q}\mu_1 - \mu_n}{(\alpha^{1/q}-1)(p+1)}\right)^{1/q}$ with $\alpha = \mu_1/\mu_n$, else $c = \mu_n^{1/q}$;
   - If some λ_i is complex, then $c = \left(\frac{\mu_1+\mu_n}{2}\right)^{1/q}$;
5. Compute $C = B^{1/q}$ by (7.5), $X = QC^{2^{k_1-k_0}}Q^T$ (or compute $C = B^{-1/q}$ by (7.4), $X = Q(C^{2^{k_1-k_0}})^{-1}Q^T$).

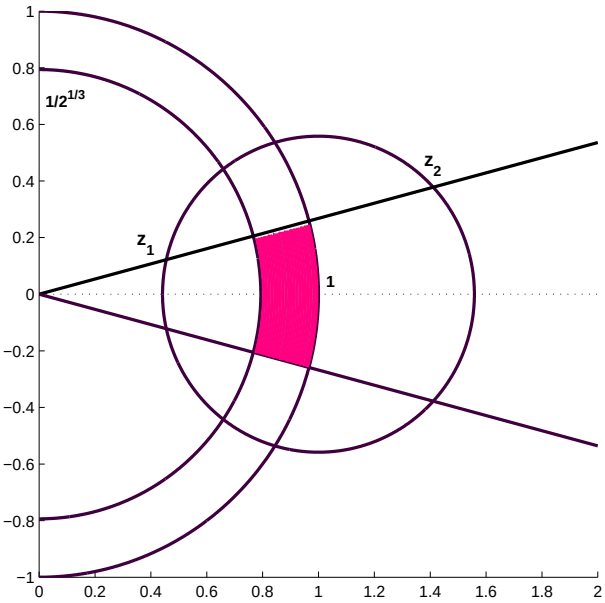

FIG. 7.1. *In gray the region $\mathcal{E}$ of the proof of Theorem 7.1 for $p = 3$.*

We show that a little extension of the region of convergence D of (6.2) allows one to implement the repeated squaring and scaling technique also for iteration (7.2). The proof is based on the proof of Theorem 2.3 of [Ian06].

THEOREM 7.1. *Iteration* (7.1) *converges for each A having eigenvalues in*

$$D_+ = \{z \in \mathbb{C} : |z| \leqslant 2, |\arg(z)| < \pi/4\}, \tag{7.6}$$

Proof. This theorem is a little extension of Theorem 2.1 of [Ian06] where it is proved that iteration (7.1) converges when A has eigenvalues in the set $\mathcal{D}$ of (6.2).

Using Proposition 2.2 of [Ian06], it is enough to prove that the set $\mathcal{E} = \{z \in \mathbb{C} : \frac{1}{2^{1/p}} \leqslant |z| < 1, |\arg(z)| < \pi/(4p)\}$ belongs to the basin of attraction of the fixed point 1 of the iteration

$$x_{k+1} = \frac{(p-1)x_k + x_k^{1-p}}{p},$$

if fact the case $|z| \geqslant 1$ is a corollary of Theorem 2.3 of [Ian06].

In Lemma 2.4 of [Ian06], it is proved that a disk centered at $z = 1$ and with radius R_p is contained in the basin of 1, where $R_p = 1 - s_p$ and s_p is the unique real solution of the equation $(2p-1)s^p - 2p\, s^{p-1} + 1 = 0$ in the interval $(0, 1)$. In Lemma 2.8 of [Ian06] it is proved that $R_p \geqslant \alpha_0/p$, for each $p > 1$, where $\alpha_0 > 1.256$.

To achieve the proof it is enough to show that the half line forming an angle of $\pi/(4p)$ with the real axis meets the circle $|z - 1| = R_p$ in two points z_1 and z_2 such that

$$r_1 < \frac{1}{\sqrt[p]{2}} < 1 < r_2,$$

where $r_1 = |z_1|$ and $r_2 = |z_2|$. That would imply that the set $\mathcal{E}$ (the dark gray set in Figure 7.1) belongs to the disk $|z-1| \leqslant R_p$ and then to the basin of attraction of the fixed point 1.

The equation that gives the two points of intersection is $|re^{\mathbf{i}\pi/(4p)} - 1| = R_p$, which can be rewritten as

$$\gamma(r) \doteq r^2 - 2r\cos(\pi/(4p)) + 1 - R_p^2 = 0.$$

The function $\gamma(r)$ is quadratic, to prove that $r_2 > 1$, observe that

$$\gamma(1) = 2 - R_p^2 - 2\cos\left(\frac{\pi}{4p}\right) \leqslant \frac{1}{p^2}\left(\frac{\pi^2}{16} - \alpha_0^2\right) < 0,$$

the inequality $r_1 < 1/\sqrt[p]{2}$ can be written as

$$\cos(\pi/(4p)) - \sqrt{\cos^2(\pi/(4p)) - 1 + R_p^2} < \frac{1}{\sqrt[p]{2}},$$

which follows from

$$\sqrt{\cos^2(\pi/(4p)) - 1 + R_p^2} \geqslant \frac{\sqrt{\alpha_0^2 - \pi/32}}{p} > 0 > \frac{\log 2}{p} \geqslant \cos(\pi/(4p)) - \frac{1}{\sqrt[p]{2}},$$

where we have used the following inequalities: $\cos^2(\pi/(4p)) - 1 \geqslant -\pi^2/(32p^2)$, $R_p^2 \geqslant \alpha_0^2/p^2$, $1/\sqrt[p]{2} > 1 - \log(2)/p$ and $\cos(\pi/(4p)) < 1$. □

Theorem 7.1 leads to a Schur-Newton algorithm using iteration (7.2).

ALGORITHM 3 (Schur-Newton algorithm using (7.2)). Given $A \in \mathbb{C}^{n\times n}$ with no nonpositive real eigenvalues, an integer $p = 2^{k_0}q$ with $k_0 \geqslant 0$ and q odd.

Use the same procedure as Algorithm 2, with $c = \left(\frac{\mu_1+\mu_n}{2}\right)^{1/q}$. In the point 2, substitute

$$\arg(\lambda_1^{1/2^{k_1}}) \in (-\pi/8, \pi/8)$$

by

$$\arg(\lambda_1^{1/2^{k_1}}) \in (-\pi/4, \pi/4)$$

From the stable version of Halley's iteration (6.8) and Theorem 6.1, we obtain another algorithm.

ALGORITHM 4 (Schur-Halley algorithm using (6.8)). Given $A \in \mathbb{C}^{n\times n}$ with no nonpositive real eigenvalues, an integer $p = 2^{k_0}q$ with $k_0 \geqslant 0$ and q odd.

Use the same procedure as Algorithm 2, with $c = \left(\frac{\mu_1+\mu_n}{2}\right)^{1/q}$. In the point 2, substitute

$$\arg(\lambda_1^{1/2^{k_1}}) \in (-\pi/8, \pi/8)$$

by

$$\arg(\lambda_1^{1/2^{k_1}}) \in (-\pi/2, \pi/2)$$

Algorithms 3 and 4 have not the disadvantages of 1, described in [GH06], i.e., a large number of steps or a possible instability in certain cases. They have the same excellent numerical behavior of 2, moreover, in some cases, they need the computation of less square roots in the preprocessing step, resulting less expensive and with a better numerical behavior.

To compare the algorithms, we use the criterion used in [GH06], considering the *relative residual*

$$\rho_A(\widetilde{X}) \doteq \frac{\|A - \widetilde{X}^p\|}{\|\widetilde{X}\| \left\| \sum_{i=0}^{p-1} \left(\widetilde{X}^{p-1-i}\right)^T \otimes \widetilde{X}^i \right\|},$$

where $\widetilde{X}$ is the computed matrix and where the norm used is the infinity norm and the algorithms are stopped when $\|N_k - I\| < 100nu$, where n is the size of A and u is the machine precision.

Algorithm 2 iteration (7.4)	Algorithm 3 iteration (7.2)	Algorithm 4 iteration (6.8)
$\rho_A(\widetilde{X}) = 3.3\text{e-}16$ $\rho_{A^{-1}}(\widetilde{X}^{-1}) = 7.4\text{e-}17$ iter=5, $k_1 = 3$	$\rho_A(\widetilde{X}) = 2.7\text{e-}16$ $\rho_{A^{-1}}(\widetilde{X}^{-1}) = 4.2\text{e-}16$ iter=5, $k_1 = 2$	$\rho_A(\widetilde{X}) = 2.8\text{e-}16$ $\rho_{A^{-1}}(\widetilde{X}^{-1}) = 4.7\text{e-}16$ iter=3, $k_1 = 2$

TABLE 7.1
Results for random nonnormal matrix.

It is computed the 5th root of a random nonnormal matrix constructed as described in [GH06] with Algorithms 2, 3 and 4. This example was used in [GH06] to show the better behavior of Algorithm 2 with respect to Algorithm 1.

In Table 7.1 we compare the results in terms of relative residual, number of steps (iter) and number of square roots in preprocessing (k_1). Observe that Algorithm 3 gives the same numerical results of Algorithm 2, with one less square roots in preprocessing. Algorithm 4 requires in general less square roots in preprocessing and a minor number of steps since it has cubic convergence, though the computational cost per step is higher than the other two. An advantage of Algorithm 4 is that it is structure-preserving.

The large region of convergence for Algorithm 4 suggests the possibility of a scaling technique which should further reduce the number of steps and the number of square roots in preprocessing, this is a topic for future research.

REFERENCES

[Bea91] Alan F. Beardon. *Iteration of rational functions*, volume 132 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1991. Complex analytic dynamical systems.

[BH03] Xavier Buff and Christian Henriksen. On König's root-finding algorithms. *Nonlinearity*, 16(3):989–1015, 2003.

[BHM05] Dario A. Bini, Nicholas J. Higham, and Beatrice Meini. Algorithms for the matrix pth root. *Numer. Algorithms*, 39(4):349–378, 2005.

[FS06] Andreas Frommer and Valeria Simoncini. Matrix Functions. Technical report, Bergische Universität Wuppertal, Fachbereich C – Mathematik und Naturwissenschaften, Wuppertal, Germany, may 2006.

[GH06] Chun-Hua Guo and Nicholas J. Higham. A Schur–Newton method for the matrix pth root and its inverse. *SIAM J. Matrix Anal. Appl.*, 28(3):788–804, 2006.

[GVL96] Gene H. Golub and Charles F. Van Loan. *Matrix computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, third edition, 1996.

[Haw02] Jane M. Hawkins. McMullen's root-finding algorithm for cubic polynomials. *Proc. Amer. Math. Soc.*, 130(9):2583–2592 (electronic), 2002.

[Hig] Nicholas J. Higham. *Functions of a Matrix: Theory and Computation*. Book in preparation.

[Hig86] Nicholas J. Higham. Newton's method for the matrix square root. *Math. Comp.*, 46(174):537–549, 1986.

[Hig97] Nicholas J. Higham. Stable iterations for the matrix square root. *Numer. Algorithms*, 15(2):227–242, 1997.

[HJ90] Roger A. Horn and Charles R. Johnson. *Matrix analysis*. Cambridge University Press, Cambridge, 1990. Corrected reprint of the 1985 original.

[HMMT04] Nicholas J. Higham, D. Steven Mackey, Niloufer Mackey, and Françoise Tisseur. Computing the polar decomposition and the matrix sign decomposition in matrix groups. *SIAM J. Matrix Anal. Appl.*, 25(4):1178–1192 (electronic), 2004.

[HMMT05] Nicholas J. Higham, D. Steven Mackey, Niloufer Mackey, and Françoise Tisseur. Functions preserving matrix groups and iterations for the matrix square root. *SIAM J. Matrix Anal. Appl.*, 26(3):849–877 (electronic), 2005.

[Hou53] Alston S. Householder. *Principles of numerical analysis*. McGraw-Hill Book Company, Inc., New York-Toronto-London, 1953.

[How83] James Lucien Howland. The sign matrix and the separation of matrix eigenvalues. *Linear Algebra Appl.*, 49:221–232, 1983.

[Ian06] Bruno Iannazzo. On the Newton method for the matrix pth root. *SIAM J. Matrix Anal. Appl.*, 28(2):503–523 (electronic), 2006.

[KL91] Charles Kenney and Alan J. Laub. Rational iterative methods for the matrix sign function. *SIAM J. Matrix Anal. Appl.*, 12(2):273–291, 1991.

[Lak98] S. Lakić. On the computation of the matrix kth root. *ZAMM Z. Angew. Math. Mech.*, 78(3):167–172, 1998.

[McM87] Curt McMullen. Families of rational maps and iterative root-finding algorithms. *Ann. of Math. (2)*, 125(3):467–493, 1987.

[Sch70] E. Schröder. Ueber unendlich viele algorithmen zur auflösung der gleichungen. *Mathematische Annalen*, 1870.

[Sch93] E. Schröder. On infinitely many algorithms for solving equations. 1993. Englis translation by G. W. Stewart.

[Smi03] Matthew I. Smith. A Schur algorithm for computing matrix pth roots. *SIAM J. Matrix Anal. Appl.*, 24(4):971–989 (electronic), 2003.