A FAMILY OF RATIONAL ITERATIONS AND ITS APPLICATION TO THE COMPUTATION OF THE MATRIX PTH ROOT

BRUNO IANNAZZO*

Abstract. Matrix fixed-point iterations $z_{n+1} = \psi(z_n)$ defined by a rational function ψ are considered. For these iterations a new proof is given that matrix convergence is essentially 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 Ernst Schröder. This characterization provides a family of iterations for the matrix pth root which preserve the structure of a group of automorphisms associated with a bilinear or a sesquilinear form. The first iteration in that family is the Halley method for which a convergence result is proved. Finally, new algorithms for the matrix pth root based on the Newton and Halley iterations are designed using the idea of the Schur–Newton method of Guo and Higham.

Key words. Halley's method, 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 which in most cases are fractals [1].

The generalization to the matrix case appears in the study of matrix equations and in the computation of matrix functions [9]. 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 a general convergence result for rational matrix iterations, then we prove some properties of specific classes of rational iterations.

General results concern the case where the iterates are rational functions of a matrix A, say $s_k(A)$. We prove that the uniform convergence of $s_k(z)$ on a compact neighborhood of the spectrum of A implies the matrix convergence, then we show that if the iteration is of the type $x_{k+1} = \varphi(x_k)$, where φ is a rational function, then 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 the matrix convergence. This extends in part a result of Higham [9, Thm. 4.15].

Concerning specific classes, we first consider the principal Padé family introduced in [17] and discussed in [8, 10, 11, 4]. We prove that the family can be obtained by the König root-finding method applied to the polynomial $x^2 - 1$, which goes back to a work of Schröder in 1870 [20]. 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 [10] about the property of a part of the principal Padé family

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

of preserving the structure of group of automorphisms associated with a bilinear or a sesquilinear form. Third, we show that the Halley method, that belongs to the König family, for the computation of the principal pth root of a matrix, preserves the structure described above and we prove a result on the convergence of that method. Finally, we show that the idea of the Schur-Newton method proposed by Guo and Higham in [6] for the inverse Newton iteration for the computation of the principal pth root of a matrix, can be applied to the direct Newton iteration and to the Halley method, providing new algorithms with good numerical properties.

We recall that the principal pth root of a matrix A having no nonpositive real eigenvalues is the unique solution X of the matrix equation $X^p - A = 0$, such that the eigenvalues of X have argument less in modulus than π/p.

The paper is organized as follows: in Section 2 we define the class of pure rational matrix iterations and we discuss their convergence; in Section 3 we show the equivalence between the principal Padé iterations and the König iterations for $x^2 - 1$; in Section 4 we generate a König family of matrix iterations preserving the structure of group of automorphisms; in Sections 5 and 6 we prove convergence results for the Newton and Halley method and we extend the idea of the Schur-Newton method of Guo and Higham to them.

2. Pure rational matrix iterations. 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 poles, so that the sequence is not necessarily well defined for each z_0. We use the notation $\varphi^{\circ k}$ to denote the kth iterate of the function φ, i.e., $\varphi^{\circ 1} = \varphi$ and $\varphi^{\circ k+1} = \varphi \circ \varphi^{\circ k}$. A fixed point z_* of (2.1) is such that $\varphi(z_*) = z_*$ and is said to be *attractive* if $|\varphi'(z_*)| < 1$. For an attractive fixed point z_*, the *basin of attraction* is the set $\mathcal{B} = \{z_0 \in \mathbb{C} : z_k \to z_*\}$; the *immediate basin* is the connected component of $\mathcal{B}$ which contains z_*.

We state an useful lemma on the basin of attraction which is a special case of Theorem 6.3.1 of [1].

LEMMA 2.1. *Let z_* be an attractive fixed point of iteration* (2.1)*. The sequence $\varphi^{\circ k}(z)$ converges locally uniformly to z_* for each z_0 belonging to the basin of z_*. In other words, z_0 has a neighborhood in which $\varphi^{\circ k}$ converges uniformly to z_*.*

Proof. Since $|\varphi'(z_*)| < 1$, there exists a closed disk D centered at z_* and such that $|\varphi(z) - z_*| \leqslant M|z - z_*|$, for a positive constant $M < 1$ and for each $z \in D$, and thus $\varphi^{\circ k}$ converges uniformly on the compact sets of D.

Let z_0 belong to the basin of attraction of z_*. There exists m such that $\varphi^{\circ m}(z_0)$ belongs to the interior of D. Since $\varphi^{\circ m}$ is continuous, there exists a compact neighborhood K of z_0 such that $\varphi^{\circ m}(z)$ is a compact set fully contained in the interior of D and thus $\varphi^{\circ k}(z)$ converges uniformly to z_* for each $z \in K$. ∎

In the matrix case, a formula like (2.1) would give an iteration of the form

$$\begin{cases} Z_0 \in \mathbb{C}^{n\times n}, \\ Z_{k+1} = \varphi(Z_k), \quad k = 0, 1, 2, \dots \end{cases} \tag{2.2}$$

where $\varphi(z)$ is a rational function and $\varphi(Z)$, where Z is a square matrix, is defined by substituting Z for z and replacing scalar numbers by multiples of the identity, and arithmetic operations by matrix operations. That procedure leads to the usual definition of function of a matrix [12, 5, 9]. We call an iteration defined by a function, as in (2.2), *pure rational matrix iteration*.

The class of pure rational matrix iterations is not suitable to approximate generic matrix functions, since, as we will explain in Remark 2.5 there hold strong conditions on the limits of such sequences.

A larger class of iterations than the pure rational matrix iterations can be studied with similar techniques. An iteration in this larger class can be written in the form

$$\begin{cases} Z_0 = p(A), \\ Z_{k+1} = \psi(Z_k, A), \quad k = 0, 1, 2, \dots \end{cases} \tag{2.3}$$

where A is a square matrix, $\psi = \psi(t, z)$ 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 $s_k(A) = \psi(s_{k-1}(A), A) = Z_k$ and $s_0(A) = p(A)$. That class contains the pure rational matrix iterations as a special case, if p is the identity function and the formula for ψ does not contain A.

Consider an iteration of the class (2.3) described above. Let Z_k be the kth iterate, so that $Z_k = s_k(A)$, with $s_k(z)$ being a rational function. Using the Jordan canonical form of A, say $M^{-1}AM = J_1 \oplus \cdots \oplus J_r$, one has $M^{-1}Z_kM = s_k(J_1) \oplus \cdots \oplus s_k(J_r)$ for each k. Therefore, by means of 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 is a Jordan block of arbitrary size for the eigenvalue λ, which will be denoted by J.

Moreover, in view of the formula for a function of a Jordan block [5, Thm 11.1.1],

$$f(J) = \begin{bmatrix} f(\lambda) & f'(\lambda) & \dots & \frac{f^{(k-1)}(\lambda)}{(k-1)!} \\ & f(\lambda) & \ddots & \vdots \\ & & \ddots & f'(\lambda) \\ \bigcirc & & & f(\lambda) \end{bmatrix}, \tag{2.4}$$

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 2.2. Consider the rational iteration $z_{k+1} = \varphi(z_k)$ where $\varphi(z) = z^2$. The sequence $\varphi^{\circ k}(1)$ converges to 1, but it fails to converge uniformly on any neighborhood of the point 1. Consider the matrix iteration $Z_{k+1} = Z_k^2$, and the starting point $Z_0 = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$; the iterates are $Z_k = \begin{bmatrix} 1 & 2^k \\ 0 & 1 \end{bmatrix}$ and the sequence fails to converge. For this iteration and Z_0 being a Jordan block of size n for the eigenvalue 1, there is matrix convergence only for $n = 1$, that is, in the scalar case. □

A sufficient condition for the convergence of the matrix sequence, given the scalar convergence, is stated in Lemma 2.3, in which the notation $\|f(z)\|_K = \sup_{x \in K} |f(x)|$ is used for a function f defined on a compact set K. This approach generalizes a proof of matrix convergence in [15]. A different approach for the matrix convergence has been used in [17] and generalized in [9, Thm. 4.15], where it is proved that the matrix convergence follows from the scalar convergence of the eigenvalues to attracting fixed points.

LEMMA 2.3. *If $s_k(z)$ is a sequence of rational functions that converges uniformly in a compact neighborhood K of λ to the function $f(z)$, then $s_k(J)$ converges to $f(J)$,*

where J is a Jordan block of arbitrary size n relative to the eigenvalue λ. Moreover, there exists a function $c = c(n)$, independent of k, such that

$$\|s_k(J) - f(J)\|_\infty \leqslant c\|s_k(z) - f(z)\|_K. \tag{2.5}$$

Proof. The function f is holomorphic on K since it is the uniform limit on a compact set of holomorphic functions.

From formula (2.4), 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 radius R, centered at λ and fully contained in K. Using the Cauchy formula, for $p = 0, 1, \dots, n-1$, it holds that

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

The previous relation provides the convergence of the sequence $s_k(J)$ to $f(J)$ since the latter term tends to zero as k tends to ∞ by the uniform convergence assumption. It provides also the proof of (2.5), since from formula (2.4) it follows that

$$\|s_k(J) - f(J)\|_\infty = \sum_{p=0}^{n-1} \left| \frac{s_k^{(p)}(\lambda)}{p!} - \frac{f^{(p)}(\lambda)}{p!} \right| \leqslant \|s_k(z) - f(z)\|_K \sum_{p=0}^{n-1} \frac{1}{R^p}. \tag{2.6}$$

□

In summary, if the sequence $s_k(z)$ converges uniformly on a compact neighborhood of the spectrum of A, then the sequence $s_k(A)$ converges and formula (2.6) can be used to provide an upper bound for the convergence of the matrix sequence. If the scalar convergence is not uniform, then the matrix iteration may fail to converge, as Example 2.2 shows.

We have turned the problem from matrix convergence to uniform scalar convergence on a compact neighborhood of the spectrum. This does not seem at first sight an advantage, but its benefit is clear in the case of pure rational iterations; in fact, Lemma 2.1 shows that if the sequence $s_k(\lambda)$ converges, for each eigenvalue λ of A, to an attractive fixed point λ_*, then the sequence $s_k(z)$ converges uniformly to λ_* on a neighborhood of λ. We have the following result.

THEOREM 2.4. *Let $Z_{k+1} = \varphi(Z_k)$ be a pure rational matrix iteration. If for each eigenvalue λ of Z_0 the scalar sequence $z_{k+1} = \varphi(z_k)$, $z_0 = \lambda$, converges to an attractive fixed point λ_*, then there exists a locally constant function $f(z)$ such that for each initial value Z in a neighborhood of Z_0 the matrix iteration converges to $f(Z)$. Moreover, $f(Z)$ is diagonalizable.*

Proof. Lemma 2.1 guarantees that the scalar iteration converges uniformly in a compact neighborhood K of spectrum of Z_0 to a locally constant function $f(z)$. Lemma 2.3 provides the matrix convergence for the Jordan blocks relative to eigenvalues belonging to the interior of K. Since the eigenvalues of a matrix are continuous functions of the entries, there exists a neighborhood V of Z_0 in the space of square matrices such that for each matrix Z of V, the eigenvalues of Z belong to the interior of K, so the matrix iteration converges to $f(Z)$.

The diagonalizability follows from the fact that f is locally constant and thus its derivatives are 0: by formula (2.4) $f(J)$ is a diagonal matrix for each Jordan block J. □

REMARK 2.5. Theorem 2.4 states that the limit of a pure rational matrix iteration is a (scalar) locally constant function of the initial value and it is diagonalizable, provided the convergence of the scalar sequence on the eigenvalues of Z_0 is to attractive fixed points (a scalar locally constant function need not be locally constant if applied to matrices, consider, for instance, the matrix sign function). A consequence is that only (scalar) locally constant functions can be the limit of a pure rational matrix iteration, which explains why in the literature the sole matrix functions computed using pure rational matrix iterations are the matrix sign function and the matrix sector function, which are (scalar) locally constant.

On the other hand, Theorem 2.4 shows that a function which is not (scalar) locally constant cannot be the limit of a pure rational matrix iteration, thus there is no hope to find, for instance, a pure rational matrix iteration converging to the matrix pth root, logarithm or exponential.

We note that Theorem 2.4 implies that a matrix function defined as the limit of a pure rational iteration is diagonalizable, which in particular gives another proof of the diagonalizability of the matrix sign function.

REMARK 2.6. A convergence result for iterations of the type (2.3) is given by Higham [9, Thm. 4.15], generalizing a result for the matrix sign function in [17, Lem. 5.1]. His result guarantees matrix convergence if the scalar eigenvalue sequences converge to attractive fixed points. When [9, Thm. 4.15], is specialized to the case of pure rational matrix iterations it gives a result similar to, but weaker than, Theorem 2.4. Theorem 2.4 together with Lemma 2.3 has the advantage of specifying the limit to which the matrix sequence converges as a matrix function, provides a bound for the matrix convergence, and can be further extended to the case $|\varphi'(\lambda^*)| = 1$ using, for instance, the Leau-Fatou theorem [1].

3. Equivalence between the König family and the principal Padé iterations family. In the paper [17] 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.

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

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

defines a family of iterations for the matrix sign function is obtained.

The iterations with $m = n-1$ and $m = n$ are globally convergent and have been called *principal Padé iterations* [9]. For these values of m and n one can define

$$g_r(x) = f_{mn}(x), \text{ for } r = m+n+1, \tag{3.1}$$

for which we have the following result [17].

THEOREM 3.1. *For the function* (3.1) *it holds that:*

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

Higham [9] noticed that these families were essentially derived by Howland [14], though for even r the iteration functions of Howland are the reciprocal of those of Kenney and Laub.

In fact, the family of principal Padé iterations is a particular case of iterations going back to Schröder in his monumental paper of 1870 [20] (an English translation is available in [21]). This family was studied by Householder [13] and many other authors, who called it König family [3] or basic family [16].

The König method of order σ, applied to the function f, is defined by the formula [3]

$$K_{f,\sigma}(z) = z + (\sigma - 1)\frac{(1/f(z))^{(\sigma-2)}}{(1/f(z))^{(\sigma-1)}}, \tag{3.2}$$

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, while for $\sigma = 3$ it 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 3.2. *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* (3.1) *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

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

□

4. Structure-preserving algorithms in the König family. It has been proved in [11, Thm. 3.13] that an iteration of the form

$$z\frac{q(z)}{\operatorname{rev} q(z)}, \tag{4.1}$$

where $\operatorname{rev} q(z) = z^d q(1/z)$, for a real polynomial $q(z)$ of degree d, preserves the structure of group of automorphisms associated with

- a bilinear form on $\mathbb{R}^n$ or $\mathbb{C}^n$;
- a sesquilinear form on $\mathbb{C}^n$.

To ease the notation we call *structure-preserving* an iteration with the form (4.1), recalling that the rational functions preserving bilinear or sesquilinear forms (fully characterized in [11, Thm. 3.13]) are more general.

The principal Padé iterations and, in view of Theorem 3.2, the $K_{2,\sigma}$ iterations, for odd σ, are iterations for the matrix sign function which are structure-preserving [11]; this is a case of a more general theorem.

THEOREM 4.1. *If $n \equiv 3 \pmod p$, then the function $K_{p,n}$, namely the König method for the polynomial $x^p - 1$, has the form*

$$z\frac{q(z^p)}{\operatorname{rev} q(z^p)}, \tag{4.2}$$

where q is a real polynomial, so in particular it is structure-preserving.

Proof. The proof is obtained by deriving a formula for the derivatives of $1/(x^p-1)$ and, from it, an explicit elementary formula for the König function from which we deduce the theorem. Let $p \geqslant 3$, the case $p = 2$ follows easily from Theorem 3.2.

Let $\omega = \cos(2\pi/p) + \mathbf{i}\sin(2\pi/p)$ and $\varphi(x) = (x^p-1)/(x-1) = \sum_{k=0}^{p-1} x^k$. 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,

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

Now,

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

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

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

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

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

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

$$x\psi_n(x)-(x^p-1)\psi_{n-1}(x) = \frac{1}{p}\sum_{k=0}^{p-1}\left(\omega^{k(n-1)}x\varphi^n(\omega^k x) - \omega^{k(n-2)}(x^p-1)\varphi^{n-1}(\omega^k x)\right)$$

$$= \frac{1}{p}\sum_{k=0}^{p-1}\left(\omega^{k(n-1)}x\varphi^n(\omega^k x) - \omega^{k(n-2)}(\omega^k x-1)\varphi^n(\omega^k x)\right) = \frac{1}{p}\sum_{k=0}^{p-1}\omega^{k(n-2)}\varphi^n(\omega^k x)$$

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

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

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

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

To conclude the proof, it is enough to prove that, for $n\equiv 3 \pmod p$, it holds that $\mu=\nu$ and $a_{\alpha_1}=a_{\beta_\mu}, a_{\alpha_2}=a_{\beta_{\mu-1}},\dots,a_{\alpha_\nu}=a_{\beta_1}$.

Let $N=\deg\varphi^n(x) = np-n$. To prove the equality $\mu=\nu$ observe that μ and ν are the number of solutions of the congruence $r\equiv 1-n \pmod p$ and $r\equiv 2-n \pmod p$, respectively, such that $0\leqslant r\leqslant N$. For $n\equiv 3 \pmod p$ there exists an integer γ such that $N=\gamma p-3$, thus the number of solutions of the two congruences $r\equiv 1-n\equiv -2 \pmod p$ and $r\equiv 2-n\equiv -1 \pmod p$ such that $0\leqslant r\leqslant N$ is the same.

Observe that since $N=np-n$, then $\beta_\mu = N+2-p$ and observe that $\varphi^n(x) = \operatorname{rev}\varphi^n(x)$, namely $a_r=a_{N-r}$ for each $r=0,1,\dots$. For $n\equiv 3 \pmod p$, it holds that $\alpha_1=p-2$ and thus $a_{\alpha_1}=a_{p-2}=a_{N+2-p}=a_{\beta_\mu}$.

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

Simplifying the common factors gives the required form for $K_{p,n}$. □

By the properties of the König method [3], 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

$$x_{k+1} = \zeta K_{p,n}(\zeta^{-1}x_k), \tag{4.3}$$

where ζ is any pth root of the nonzero scalar a, for $x_0=\zeta z_0$, is such that $x_k=\zeta z_k$ and thus converges locally to the roots of $x^p-a=0$.

Iteration (4.3) does not seem effective for computing the pth roots of a, since it uses ζ, but for $n\equiv 3 \pmod p$, in view of Theorem 4.1, the iteration for x_k has the form

$$x_{k+1} = x_k\frac{q(a^{-1}x_k^p)}{\operatorname{rev} q(a^{-1}x_k^p)} = x_k\frac{\widehat{q}(x_k^p)}{\operatorname{rev}\widehat{q}(x_k^p)}, \tag{4.4}$$

where $\widehat{q}$ is obtained multiplying q by a suitable power of a. In this way, an effective iteration is obtained to approximate with a high precision the pth roots of a given complex number.

A difficulty in the use of iteration (4.4) is the global convergence. We will not investigate further the global convergence of (4.4), but in Section 5 we will give a convergence proof for the case $n = 3$ which is a structure-preserving iteration for each p, in view of Theorem 4.1.

REMARK 4.2. Theorem 4.1 has a perhaps surprising application to the theory of root-finding algorithms. Following McMullen [19], 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 cubics. 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 [19] a generally convergent algorithm for cubic polynomials and has proved that there does not exist a generally convergent algorithm for polynomials 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 (4.2) [7]. Thus, Theorem 4.1 could be used to construct generally convergent algorithms for cubic polynomial of arbitrarily high order of convergence.

5. Nice properties of the Halley method. The König method of order 3 is the so-called Halley method which, for the equation $x^p - 1 = 0$, is

$$x_{k+1} = x_k \frac{(p-1)x_k^p + (p+1)}{(p+1)x_k^p + (p-1)}, \quad x_0 \in \mathbb{C}. \tag{5.1}$$

Here we considered a matrix generalization of the Halley method for computing the principal pth root of a matrix A.

A very nice feature of the Halley method for the equation $x^p - 1 = 0$ is that the basin of attraction for the fixed point 1 is somewhat *nicer* than the one of the Newton method (see Figure 5.1 for a comparison in the case $p = 4$). It has been proved [15] that for the Newton method applied to the equation $x^p - 1 = 0$ the basin of attraction for the fixed point 1 contains the set

$$\mathcal{T}_{2p} = \{z \in \mathbb{C} \setminus \{0\} \,:\, -\pi/(2p) < \arg(z) < \pi/(2p), |z| \geqslant 1\}, \tag{5.2}$$

while for the Halley method there holds the following result.

THEOREM 5.1. *The immediate basin of attraction for the fixed point* 1 *of the rational iteration* (5.1) *contains the sector*

$$\mathcal{S}_{2p} = \{z \in \mathbb{C} \setminus \{0\} \,:\, -\pi/(2p) < \arg(z) < \pi/(2p)\}. \tag{5.3}$$

Proof. Let us define

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

iteration (5.1) can be written as $z_{k+1} = z_k\varphi(z_k)$. The sector $\mathcal{S}_{2p}$ contains the fixed point $z = 1$, is an open connected set and, by Lemma 5.2, 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,

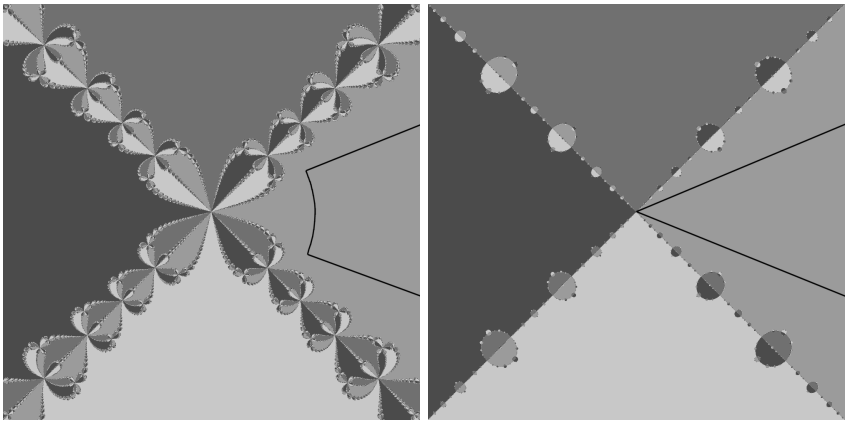

FIG. 5.1. *Comparison of the basins of attraction for the Newton method (left) and the Halley method (right) for the equation* $x^4 - 1$ *in the set* $[-2, 2] \times [-2, 2] \subset \mathbb{C}$. *The sets* $\mathcal{T}_{2p}$ *of* (5.2) *and* $\mathcal{S}_{2p}$ *of* (5.3) *are highlighted.*

given a rational iteration $x_{k+1} = \psi(x_k)$ of degree greater than 1, any connected open set $\mathcal{U}$ such that $\psi(\mathcal{U}) \subset \mathcal{U}$ and containing only a fixed point z_* belongs to the immediate basin of z_* (compare [1, Thm. 4.2.5]). □

LEMMA 5.2. *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

$$-2\arg(z) < \arg(\varphi(z)) < 0. \tag{5.4}$$

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 (5.4) it is obtained the equivalent

$$-\frac{\sin(2\vartheta)}{\cos(2\vartheta)} < \frac{-4pr^p\sin(p\vartheta)}{(p^2-1)(r^{2p}+1) + 2(p^2+1)r^p\cos(p\vartheta)} < 0. \tag{5.5}$$

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

$$(p^2-1)\sin(2\vartheta)r^{2p} + 2\big((p^2+1)\cos(p\vartheta)\sin(2\vartheta) - 2p\sin(p\vartheta)\cos(2\vartheta)\big)r^p + (p^2-1)\sin(2\vartheta) > 0, \tag{5.6}$$

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 $0 < \vartheta < \pi/p$ and $\sin((p-2)\vartheta) = \sin(p\vartheta)\cos(2\vartheta) - \sin(2\vartheta)\cos(p\vartheta)$, one can see that

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

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

It is worth giving a corollary of Theorem 5.1 which could be used for the computation of the scalar p-th root.

COROLLARY 5.3. *Consider the Halley method for the equation* $x^p - a = 0$,

$$x_{k+1} = x_k \frac{(p-1)x_k^p + (p+1)a}{(p+1)x_k^p + (p-1)a}, \quad x_0 \in \mathbb{C}. \tag{5.7}$$

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

Theorem 2.4 guarantees the convergence of the pure matrix iteration

$$Y_{k+1} = Y_k((p-1)Y_k^p + (p+1)I)((p+1)Y_k^p + (p-1)I)^{-1}, \tag{5.8}$$

to the identity matrix I, for each Y_0 having eigenvalues in $\mathcal{S}_{2p}$; in particular, for $Y_0 = A^{-1/p}$ where A has eigenvalues in the open right half complex plane, which will be denoted by $\mathbb{C}_>$. Iteration (5.8) is strictly related to

$$X_{k+1} = X_k((p-1)X_k^p + (p+1)A)((p+1)X_k^p + (p-1)A)^{-1}; \tag{5.9}$$

in fact, if A has eigenvalues in $\mathbb{C}_>$, $Y_0 = A^{-1/p}$ and $X_0 = I$, then it can be shown that $X_k = Y_k A^{1/p}$ for each k (the proof follows by an induction argument and using the fact that X_k and Y_k are functions of A, so commute with A and $A^{1/p}$).

COROLLARY 5.4. *The sequence* X_k *obtained by iteration* (5.9) *with* $X_0 = I$ *converges to* $A^{1/p}$ *for each* A *having eigenvalues in* $\mathbb{C}_>$.

Moreover, for what we have proved in Section 4, iteration (5.8) is structure-preserving. If A belongs to a group of automorphisms as in Section 4, so does $A^{1/p}$, thus, each of the iterates obtained by (5.9) belongs to that group.

Iteration (5.9) cannot be used directly to approximate the principal pth root. In fact, using the same idea as in [15], one can prove that iteration (5.9) is *not stable 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 overridden using another algorithm which provides the same sequence but which is stable in a neighborhood of $A^{1/p}$, for instance

$$\begin{cases} X_0 = I, \quad N_0 = A, \\ X_{k+1} = X_k((p+1)I + (p-1)N_k)^{-1}((p-1)I + (p+1)N_k), \\ N_{k+1} = N_k\left(((p+1)I + (p-1)N_k)^{-1}((p-1)I + (p+1)N_k)\right)^{-p}. \end{cases} \tag{5.10}$$

where $N_k \to I$ and $X_k \to A^{1/p}$. If the pth power is computed using the binary powering technique [5, Alg. 11.2.2], the computational cost of iteration (5.10) is $2(5 + \vartheta\log_2 p)n^3$ arithmetic operations (ops) per step, where $1 \leqslant \vartheta \leqslant 2$.

6. New algorithms for the matrix pth root. A family of iterations for computing the principal pth root of a matrix A is

$$X_{k+1} = \frac{(p-1)X_k + AX_k^{1-p}}{p}, \tag{6.1}$$

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 [22], this is the reason why iteration (6.1) is referred, somehow improperly, as the Newton method.

In [22] it is proved that this iteration is not stable in a neighborhood of $A^{1/p}$. A stable variant, for $X_0 = I$,

$$\begin{cases} Y_0 = I, \quad N_0 = A, \\ Y_{k+1} = Y_k \left(\dfrac{(p-1)I + N_k}{p} \right), \\ N_{k+1} = \left(\dfrac{(p-1)I + N_k}{p} \right)^{-p} N_k, \end{cases} \tag{6.2}$$

has been proposed in [15], 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

$$\mathcal{D} = \{z \in \mathbb{C} : \operatorname{Re} z > 0, |z| \leqslant 1\}. \tag{6.3}$$

This leads to an algorithm for computing the principal pth root.

ALGORITHM 1 (A Newton method for $A^{1/p}$ [15]). 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 (6.3);
3. By means of iteration (6.2)
   - 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 $X = \left(S\|B\|^{1/p}\right)^2$.

Iteration (6.2) of Algorithm 1 has a computational cost of $2(3+\vartheta \log_2 p)n^3$ ops per step, where $1 \leqslant \vartheta \leqslant 2$. The initial square root can be obtained by forming the Schur decomposition of A, without affecting the complexity order with respect to p. An observation of Guo and Higham is that the Schur decomposition gives the eigenvalues of A and that information is not exploited in Algorithm 1.

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 could reduce the number of steps needed for the expensive iteration (6.2) (or other similar) to verify a suitable stopping criterion. 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 further reduce the number of steps needed for convergence.

The *Schur-Newton method*, an algorithm of Guo and Higham [6], is based on these ideas. The algorithm does not use iteration (6.2) but an iteration which generalizes the scalar Newton method for the equation $x^{-p} - a = 0$. The iteration, introduced in

[2], is

$$X_{k+1} = \frac{1}{p}\left((p+1)X_k - X_k^{p+1}A\right), \quad X_0 = I, \tag{6.4}$$

which converges to $A^{-1/p}$, and for which in [6] is constructed a 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 (6.4) can be obtained a stable iteration [15, 18, 6]

$$\begin{cases} Y_0 = \dfrac{1}{c}I, \quad N_0 = \dfrac{1}{c^p}A, \\ Y_{k+1} = Y_k\left(\dfrac{(p+1)I - N_k}{p}\right), \\ N_{k+1} = \left(\dfrac{(p+1)I - N_k}{p}\right)^p N_k, \end{cases} \tag{6.5}$$

such that $Y_k \to A^{-1/p}$ and $N_k \to I$. Setting $X_k = Y_k^{-1}$ gives the iteration [6]

$$\begin{cases} X_0 = cI, \quad N_0 = \dfrac{1}{c^p}A, \\ X_{k+1} = \left(\dfrac{(p+1)I - N_k}{p}\right)^{-1} X_k, \\ N_{k+1} = \left(\dfrac{(p+1)I - N_k}{p}\right)^p N_k, \end{cases} \tag{6.6}$$

for which $X_k \to A^{1/p}$. The computational costs of iterations (6.5) and (6.6) are $2(2 + \vartheta \log_2 p)n^3$ and $2(3 + \vartheta \log_2 p)n^3$ ops per step, respectively, where $1 \leqslant \vartheta \leqslant 2$.

ALGORITHM 2 (Schur-Newton algorithm for $A^{1/p}$ using (6.5) and (6.6) [6]). 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 $\arg(\lambda_i^{1/2^{k_1}}) \in (-\pi/8, \pi/8)$ for each i 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 taking the square root k_1 times; if $q = 1$, then $X = QBQ^T$, 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 (6.6), $X = QC^{2^{k_1-k_0}}Q^T$ (or compute $C = B^{-1/q}$ by (6.5), $X = Q(C^{2^{k_1-k_0}})^{-1}Q^T$).

The initial square roots computation, in certain cases, may reduce dramatically the number of steps needed by the iteration, but each square root in preprocessing corresponds to a squaring at the final step of the algorithm. The cost of a square root and a squaring is less than the cost of one step of the iteration, but a large number of initial square roots may result in a waste of computation if there is no saving in the number of iteration steps.

A little extension of the region of convergence D of (6.3) allows one to use the ideas of Algorithm 2 also for iteration (6.2). The proof will be given in Section 6.1 and is based on the proof of Theorem 2.3 of [15].

THEOREM 6.1. *The immediate basin of attraction for the fixed point 1 of the iteration*

$$x_{k+1} = \frac{(p-1)x_k + x_k^{1-p}}{p}, \tag{6.7}$$

contains the set

$$\mathcal{E} = \{z \in \mathbb{C} : |z| \geqslant \frac{1}{2^{1/p}}, |\arg(z)| < \pi/(4p)\}.$$

Observe that iteration (6.1) with $X_0 = I$ converges to $A^{1/p}$ if and only if the iteration

$$X_{k+1} = \frac{(p-1)X_k + X_k^{1-p}}{p}, \quad X_0 = A^{-1/p}, \tag{6.8}$$

converges to the identity matrix. This fact and Theorem 2.4 give the following result.

COROLLARY 6.2. *Iteration* (6.1) *converges for each A having eigenvalues in*

$$D_+ = \{z \in \mathbb{C} : |z| \leqslant 2, |\arg(z)| < \pi/4\}. \tag{6.9}$$

Corollary 6.2 leads to an analog of Algorithm 2 using iteration (6.2).

ALGORITHM 3 (Schur-Newton algorithm using (6.2)). 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.

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 there exists a positive number s such that for each eigenvalue λ of A, $s\lambda^{1/2^{k_1}} \in \mathcal{D}$, where $\mathcal{D}$ is the disk of center 6/5 and radius 3/4.
3. Compute $B = R^{1/2^{k_1}}$ by taking the square root k_1 times; if $q = 1$, then $X = QBQ^T$, else continue;
4. Compute $C = (B/s)^{1/q}$ by (6.2), $X = Q(Cs^{1/q})^{2^{k_1-k_0}}Q^T$.

The convergence of Algorithm 3 is guaranteed by Corollary 6.2, in fact iteration (6.2) is applied to a matrix having eigenvalues in the set $\mathcal{D}$ which is a subset of D_+ of (6.9). The set $\mathcal{D}$ is chosen heuristically in order to need at most 5 steps of Newton iteration in the scalar case.

Step 2 of Algorithm 3 can be done in an inexpensive way. For $m \geqslant k_0$ and for each eigenvalue λ of A, one looks for an interval $[t_1(\lambda), t_2(\lambda)]$ such that $t_1(\lambda) > 0$ and $t\lambda^{1/2^m}$ lies into $\mathcal{D}$ for $t \in [t_1(\lambda), t_2(\lambda)]$; if such an interval exists for each λ and the intersection is not void, then s can be any point of the intersection, else increase m.

In Figure 6.1 we have constructed experimentally the *level sets of convergence* for iterations (6.1) and (6.4) applied to scalar numbers. Given the tolerance $\varepsilon = 10^{-15}$, a point x_0 of the region $[-1, 5] \times [-3, 3]$ of the complex plane has been colored by a tonality of grey if convergence up to ε occurs in less than 10 steps. Each tonality of grey corresponds to a different number of iterations needed: the lighter one corresponds to the points for which convergence up to ε occurs in 9 steps. The black contour encloses the sets in which the eigenvalues of the matrix preprocessed by Algorithm 3 and 2 lie;

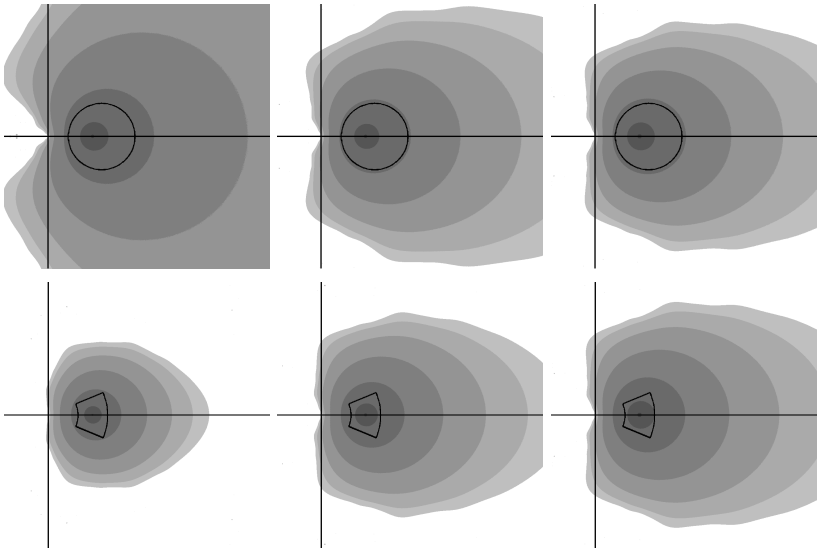

FIG. 6.1. *Comparison of the* level sets of convergence *for the Newton method (first row) and the inverse Newton method (second row) for* $p = 4$ *(first column),* $p = 20$ *(second column) and* $p = 200$ *(last column). The black contour encloses the regions in which lie the eigenvalues of the preprocessed matrices to which is applied the iterative step of Algorithms 3 (first row) and 2 (second row).*

observe that in the examples in Figure 6.1 the scalar iteration with an initial value inside the bordered regions needs at most 5 iterations. The expected number for the matrix iteration is the same, unless the matrix is nondiagonalizable.

In practice, due to the larger level sets of convergence (see Figure 6.1) Algorithm 3 is likely to obtain the same number of iteration steps as Algorithm 2 with a slightly milder condition on step 2, which could save a couple of square roots in preprocessing.

From the stable version of Halley's iteration (5.10) and Corollary 5.4, we obtain another algorithm.

ALGORITHM 4 (Schur-Halley algorithm using (5.10)). 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 there exists a positive number s such that for each eigenvalue λ of A, $s\lambda^{1/2^{k_1}} \in \mathcal{D}$, where $\mathcal{D}$ is the disk of center 8/5 and radius 1.
3. Compute $B = R^{1/2^{k_1}}$ by taking the square root k_1 times; if $q = 1$, then $X = QBQ^T$, else continue;
4. Compute $C = (B/s)^{1/q}$ by (5.10), $X = Q(Cs^{1/q})^{2^{k_1-k_0}}Q^T$.

The convergence of Algorithm 4 is guaranteed by Corollary 5.4, in fact iteration (5.10) is applied to a matrix having eigenvalues in the set $\mathbb{C}_>$.

Once again, the choice of $\mathcal{D}$ is heuristic and it is based on the observation of the experimental regions of convergence. With this preprocessing the iteration usually needs 3 steps to converge.

Algorithms 3 and 4 have not the disadvantages of Algorithm 1, described in [6], i.e., a large number of steps or a possible instability in certain cases. They have the

same excellent numerical behavior of Algorithm 2, moreover, in most cases, they can save some square root in preprocessing.

To compare the algorithms, we use the criterion used in [6], 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.

TABLE 6.1
Results for the 5th root of a random nonnormal matrix.

Algorithm 2 iteration (6.5)	Algorithm 3 iteration (6.2)	Algorithm 4 iteration (5.10)
$\rho_A(\widetilde{X}) = 3.3\text{e-}16$ $\rho_{A^{-1}}(\widetilde{X}^{-1}) = 7.4\text{e-}17$ 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 6.2
Results for the 15th root of a 3-by-3 matrix A with real eigenvalues and condition number $\kappa_2(A) \approx 10^{10}$.

Algorithm 2 iteration (6.5)	Algorithm 3 iteration (6.2)	Algorithm 4 iteration (5.10)
err = 2.7e-8 $\rho_A(\widetilde{X}) = 5.0\text{e-}17$ iter=5, $k_1 = 5$	err = 2.7e-8 $\rho_A(\widetilde{X}) = 8.1\text{e-}18$ iter=5, $k_1 = 4$	err = 2.7e-8 $\rho_A(\widetilde{X}) = 1.5\text{e-}17$ iter=3, $k_1 = 4$

As first test, it is computed the 5th root of a random nonnormal matrix constructed as described in [6] with Algorithms 2, 3 and 4. This example was used in [6] to show the better behavior of Algorithm 2 with respect to Algorithm 1. In Table 6 we compare the results in terms of relative residual, number of steps (iter) and number of square roots in preprocessing (k_1).

A second test is made considering a the nonnormal matrix

$$S = \begin{bmatrix} -1 & -2 & 2 \\ -4 & -6 & 6 \\ -4 & -16 & 13 \end{bmatrix}$$

whose eigenvalues are $\{1, 2, 3\}$ and computing the 15th root of $A \doteq S^{15}$, which is formed exactly. The condition number $\kappa_2(A) = \|A\|_2\|A^{-1}\|_2$ of the matrix A is about 10^{10}. In Table 6 the algorithms are compared in terms of the relative residual and the relative error of the computed solution $\widetilde{X}$, namely $\text{err} = \|\widetilde{X} - S\|/\|S\|$, where the Frobenius norm is used.

Observe that Algorithm 3 gives the same numerical results of Algorithm 2, with fewer square roots in preprocessing. Algorithm 4 requires in general fewer 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.

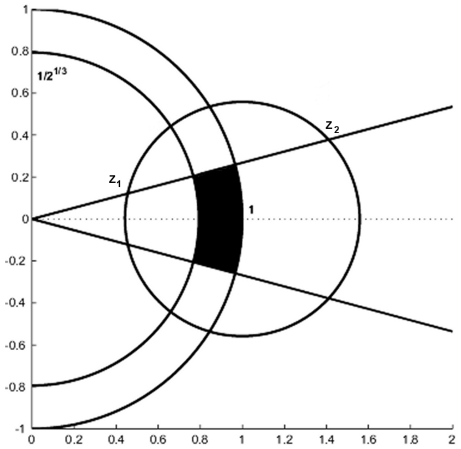

FIG. 6.2. *In black the region $\mathcal{E}$ of the proof of Theorem 6.1 for $p = 3$.*

6.1. Proof of Theorem 6.1. It is enough to prove that the set $\mathcal{E} \cap \{|z| < 1\}$ belongs to the immediate basin of attraction, in fact the case $|z| \geqslant 1$ is a corollary of Theorem 2.3 of [15].

In Lemma 2.4 of [15], 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 [15] 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 black set in Figure 6.2) 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

$$\sqrt{\cos^2(\pi/(4p)) - 1 + R_p^2} \geqslant \frac{\sqrt{\alpha_0^2 - \pi^2/16}}{p} > 0 > \frac{\log 2}{p} \geqslant \cos(\pi/(4p)) - \frac{1}{\sqrt[p]{2}},$$

where we have used the following inequalities: $\cos^2(\pi/(4p)) - 1 \geqslant -\pi^2/(16p^2)$, $R_p^2 \geqslant \alpha_0^2/p^2$, $1/\sqrt[p]{2} > 1 - \log(2)/p$ and $\cos(\pi/(4p)) < 1$. □

Acknowledgment. I would like to thank Prof. Dario A. Bini, Prof. Nicholas J. Higham and an anonymous referee whose pertinent and detailed suggestions improved considerably the presentation and the correctness of the paper.

REFERENCES

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

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

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

[4] Andreas Frommer and Valeria Simoncini. Matrix Functions. In W. Schilders and H. A. Van der Vorst, editors, *Model Order Reduction: Theory, Research Aspects and Applications*. Springer-Verlag, Berlin, 2008. To appear.

[5] 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.

[6] 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.

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

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

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

[10] 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, 2004.

[11] 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, 2005.

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

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

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

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

[16] Yi Jin and Bahman Kalantari. Symmetric functions and root-finding algorithms. *Adv. in Appl. Math.*, 34(1):156–174, 2005.

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

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

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

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

[21] Ernst Schröder. On infinitely many algorithms for solving equations. Technical report, Technoical Report TR-92-121. Department of Computer Science, University of Maryland, College Park, MD, USA, November 1992. Translated by G. W. Stewart.

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