ANNALES
POLONICI MATHEMATICI
LII (1991)

Newton-like algorithms for kth root calculation

by Marek Kuczma (Katowice) and Halina Świątak (Kraków)

Abstract. The algorithm

$$(*)\qquad a_{n+1} = pNa_n^{1-k}+(1-p)a_n, \quad n\in \mathbf{N},$$

for numerical calculation of $N^{1/k}$ (N is here a positive number, $k \geqslant 2$ an integer) is considered. The convergence of $(*)$ is investigated, and the speed of convergence is estimated, depending on the value of the parameter $p\in(0, 1)$. For $p = 1/k$, algorithm $(*)$ becomes the well-known Newton method of solving the equation $x^k-N = 0$.

As a matter of fact, $(*)$ yields the iterative sequence of the function $f: \mathbf{R}^+ \to \mathbf{R}^+$,

$$f(x) = pNx^{1-k}+(1-p)x, \quad x\in \mathbf{R}^+,$$

and the behaviour of $(*)$ is derived from the properties of f. Numerical aspects of the results obtained are also discussed.

In this paper we study the algorithm

$$(1)\qquad a_{n+1} = p\frac{N}{a_n^{k-1}}+qa_n, \quad n\in \mathbf{N},$$

for numerical calculation of $\sqrt[k]{N}$ on computers. Here $k \geqslant 2$ is a fixed integer, N is a fixed positive number (not necessarily an integer), p and q are fixed positive numbers adding up to 1:

$$(2)\qquad p+q = 1,$$

and the initial term a_1 of the sequence $\{a_n\}$ is taken arbitrarily from $\mathbf{R}^+ = (0, \infty)$. We are going to investigate the convergence and the rate of convergence of sequences $\{a_n\}$ defined by (1) depending on the values of k and p. Some numerical implications are discussed at the end of the paper.

Relation (1) generalizes Newton's algorithm for square roots calculation (cf. [4], [1])

$$a_{n+1} = \tfrac{1}{2}\left(\frac{N}{a_n}+a_n\right), \quad n\in \mathbf{N},$$

or, with arbitrary integer $k \geqslant 2$,

(3) $$a_{n+1} = \frac{1}{k}\left(\frac{N}{a_n^{k-1}} + (k-1)a_n\right), \quad n \in \boldsymbol{N},$$

which results from *Newton–Raphson's method* (or shortly *Newton's method*; cf. [1], [2])

(4) $$a_{n+1} = a_n - \frac{F(a_n)}{F'(a_n)}, \quad n \in \boldsymbol{N},$$

for numerical solution of the equation $F(x) = 0$ on taking $F(x) = x^2 - N$ or $F(x) = x^k - N$, respectively.

Relation (1) can equivalently be written as

(5) $$a_{n+1} = f(a_n), \quad n \in \boldsymbol{N},$$

or (f^n denotes the nth iterate of the function f)

(6) $$a_{n+1} = f^n(a_1), \quad n \in \boldsymbol{N}_0 = \boldsymbol{N} \cup \{0\},$$

where the function $f\colon \boldsymbol{R}^+ \to \boldsymbol{R}^+$ is given by

(7) $$f(x) = pNx^{1-k} + qx, \quad x \in \boldsymbol{R}^+$$

In the sequel we shall make use of the following definitions.

DEFINITION 1. Let I be a real interval and let ξ be a point in the closure of I. $S_\xi^0[I]$ denotes the class of continuous functions $f\colon I \to \boldsymbol{R}$ such that

(8) $$0 < \frac{f(x) - \xi}{x - \xi} < 1, \quad x \in I,\ x \neq \xi.$$

Relation (8) says that the graph of the function $y = f(x)$ lies between the straight lines $y = x$ and $y = \xi$.

DEFINITION 2. Let $\{\alpha_n\}$ and $\{\beta_n\}$ be two sequences of positive numbers. We write

$$\alpha_n \sim \beta_n, \quad n \to \infty,$$

iff the limit $\lim_{n\to\infty} \alpha_n/\beta_n$ exists, is finite and positive.

The following simple lemma (cf. [5], Theorem 0.4) is of fundamental importance.

LEMMA 1. *Let I and ξ be as in Definition* 1 *and let f be a function defined on I. If $f \in S_\xi^0[I]$, then for every $x \in I$ we have $\lim_{n\to\infty} f^n(x) = \xi$. If, moreover, $x \neq \xi$, then the sequence $\{f^n(x)\}$ is strictly monotonic.*

In what follows f always denotes the function (7). We introduce also a few symbols whose meaning will not be explained again in the sequel.

$r = N^{1/k}$ is the only fixed point of f in $\boldsymbol{R}^+$:

$$f(r) = r. \tag{9}$$

Thus $f(x) \neq x$ in $(0, r) \cup (r, \infty)$ and since evidently

$$\lim_{x \to 0} f(x) = \infty, \quad \lim_{x \to \infty} x^{-1} f(x) = q < 1, \tag{10}$$

we actually have

$$f(x) > x \quad \text{in } (0, r), \quad f(x) < x \quad \text{in } (r, \infty). \tag{11}$$

$s = f'(r) = 1 - pk$ is the multiplier of the fixed point $x = r$ of f.

$z = [(k-1)pN/q]^{1/k}$ is the only point in $\boldsymbol{R}^+$ at which the derivative of f:

$$f'(x) = (1-k)pNx^{-k} + q, \quad x \in \boldsymbol{R}^+, \tag{12}$$

vanishes. The derivative (12) is strictly increasing in $\boldsymbol{R}^+$ and when x varies from 0 to $+\infty$, the function $f'(x)$ grows from $-\infty$ to $q > 0$ so that we have

$$f'(x) < 0 \quad \text{in } (0, z), \quad f'(x) > 0 \quad \text{in } (z, \infty). \tag{13}$$

Consequently f is convex and attains its minimum at z.

$u = f(z)$ is the minimal value of f in $\boldsymbol{R}^+$:

$$f(x) \geqslant u, \quad x \in \boldsymbol{R}^+ \tag{14}$$

In particular, as a consequence of (5) and (14)

$$a_n \geqslant u, \quad n \geqslant 2, \tag{15}$$

for any sequence $\{a_n\}$ of positive numbers fulfilling (1).

$M \subset \boldsymbol{R}^+$ denotes the set

$$M = \bigcup_{m=0}^{\infty} \{x \in \boldsymbol{R}^+ : f^m(x) = r\}. \tag{16}$$

M is the set of all iterative predecessors of r and thus in view of (9), M is the orbit of r under f (cf. [5]–[7]). We always have $r \in M$ and when $pk = 1$ this is the only point of M. On the other hand, if $pk > 1$, then M is countably infinite.

The following lemma is an immediate consequence of the properties of orbits (cf. [7]), but it can also easily be established directly.

LEMMA 2. *Let $\{a_n\}$ be a sequence of positive numbers fulfilling* (1). *Then either*

$$a_n \in M, \quad n \in N, \tag{17}$$

or

$$a_n \notin M, \quad n \in N. \tag{18}$$

Next we prove:

Lemma 3. *If $k \leqslant 4$ and $pk \leqslant 2$, then*

(19) $$f^2(x) \neq x, \quad x \in \boldsymbol{R}^+,\ x \neq r.$$

Proof. According to (7), $f(x) = (pN+qx^k)/x^{k-1}$, whence

$$f^2(x) = \frac{pN+q(f(x))^k}{(f(x))^{k-1}} = \frac{pNx^{k(k-1)}+q(pN+qx^k)^k}{x^{k-1}(pN+qx^k)^{k-1}}.$$

Thus the equation $f^2(x) = x$ may be written as

(20) $$pNx^{k(k-1)}+q(pN+qx^k)^k = x^k(pN+qx^k)^{k-1},$$

or with $y = x^k$,

(21) $$pNy^{k-1}+q(pN+qy)^k = y(pN+qy)^{k-1},$$

that is,

(22) $$yw^{k-1}-qw^k-pNy^{k-1} = 0,$$

where we have put for short

(23) $$w = pN+qy.$$

The thing to show is that (20) has no positive solution except $x = r$, or what amounts to the same, that (21) has no positive solution except $y = N$. We will deal with (21) in the form (22) with (23); we aim at showing that this system of equations has no solution such that $y > 0$ except $y = w = N$. We have by (2) and (23)

$$y-qw = (1-q^2)y-pqN = p(1+q)y-pqN = p[y+q(y-N)],$$

so that (22) can be written as

(24) $$yw^{k-1}+q(y-N)w^{k-1}-Ny^{k-1} = 0.$$

Since by (2) and (23), $w-y = -p(y-N)$, adding and subtracting in (24) the term Nw^{k-1} and making use of the identity

$$w^{k-1}-y^{k-1} = (w-y)\sum_{i=0}^{k-2} w^i y^{k-2-i}$$

we obtain

$$(y-N)\left[w^{k-1}+qw^{k-1}-pN\sum_{i=0}^{k-2} w^i y^{k-2-i}\right] = 0.$$

So we have to check whether the system of equations (23) and

(25) $$(q+1)w^{k-1}-pN\sum_{i=0}^{k-2} w^i y^{k-2-i} = 0$$

has a solution such that $y > 0$ apart from $y = w = N$.

1. $k = 2$. Then (25) becomes $(q+1)w-pN = 0$, whence $w = pN/(q+1)$ and by (23), $y = -pN/(q+1) < 0$.

2. $k = 3$. Then (25) becomes $(q+1)w^2-pN(w+y) = 0$, i.e., by (23),

$$q(q+1)w^2-(q+1)pNw+p^2N^2 = 0. \tag{26}$$

The discriminant of (26) is

$$\Delta = (q+1)^2p^2N^2-4q(q+1)p^2N^2 = (q+1)p^2N^2(1-3q).$$

If $kp = 3p < 2$, then by (2) we have $3q > 1$ so that $\Delta < 0$ and (26) has no real solution. If $kp = 3p = 2$, then $3q = 1$, $\Delta = 0$, and (26) has the double root $w = N$. Then by (23) also $y = N$.

3. $k = 4$. Then (25) becomes

$$(q+1)w^3-pN(y^2+yw+w^2) = 0. \tag{27}$$

By (23) we have $(q+1)w-pN = q(w+y)$ and thus (27) turns into

$$(w+y)(qw^2-pNy) = 0.$$

For $y > 0$ we have $w+y > 0$ and thus we are led to investigate the equation

$$q^2w^2-pNw+p^2N^2 = 0. \tag{28}$$

Its discriminant is

$$\Delta = p^2N^2-4q^2p^2N^2 = p^2N^2(1+2q)(1-2q).$$

If $kp = 4p < 2$, then $2q > 1$, $\Delta < 0$, and (28) has no real solution. If $kp = 4p = 2$, then $2q = 1$ and (28) has the double root $w = N$. By (23) also $y = N$.

We see that in all considered cases the system of equations (23) and (25) has no solution such that $y > 0$ except $y = w = N$, which completes the proof of the lemma.

Remark 1. We conjecture that (19) is true for all positive integers $k \geqslant 2$ whenever $pk \leqslant 2$, but we have been unable to prove this in full generality. Anyhow, the cases most important and most often encountered in practice are just $k = 2$ and $k = 3$ (square and cubic roots).

Now we pass to the investigation of the behaviour of sequences $\{a_n\} \subset \boldsymbol{R}^+$ fulfilling (1). According to Lemma 2, such sequences have to satisfy either (17) or (18). In the former case, the situation is trivial.

THEOREM 1. *Let $\{a_n\}$ be a sequence of positive numbers fulfilling* (1) *and* (17). *Then $\{a_n\}$ is stationary: there exists an integer $m \geqslant 0$ such that*

$$a_n = r, \quad n > m. \tag{29}$$

In particular,

$$\lim_{n\to\infty} a_n = r. \tag{30}$$

Proof. By (17) we have $a_1 \in M$, which means according to (16) and (6) that there exists an integer $m \geqslant 0$ such that

(31) $$a_{m+1} = r.$$

Relation (29) follows from (31) by induction in view of (5) and (9). Relation (30) is a trivial consequence of (29).

In case (18) the behaviour of $\{a_n\}$ depends further on the multiplier s of r and thus, in fact, on pk.

THEOREM 2. *Let $\{a_n\}$ be a sequence of positive numbers fulfilling* (1) *and* (18). *If $0 < pk < 1$, then for $n \geqslant 2$ the sequence $\{a_n\}$ is strictly monotonic and* (30) *holds true. Moreover,*

(32) $$|a_n - r| \sim s^n, \quad n \to \infty.$$

Proof. Since $f'(r) = 1 - pk > 0$, we have $z < r$ by (13), which in turn implies in view of (11), (14), (9) and the strict monotonicity of f on (z, ∞) (cf. (13)) that

$$z < f(z) = u < f(r) = r.$$

Hence, again by the strict monotonicity of f on (z, ∞),

(33) $$f(x) < r \quad \text{in } [u, r), \quad f(x) > r \quad \text{in } (r, \infty).$$

Relations (11) and (33) show that (8) is fulfilled in $[u, \infty)$, that is, $f \in S_r^0[[u, \infty)]$, whereas (15) and (18) imply that $a_2 \in [u, \infty)$, $a_2 \neq r$. Disregarding the first term a_1 of the sequence $\{a_n\}$, we may consider $\{a_n\}$ as the iterative sequence $\{f^n(a_2)\}$. The strict monotonicity of $\{a_n\}$ for $n \geqslant 2$ and relation (30) result now from Lemma 1 and the asymptotic condition (32) is a consequence of a theorem of Thron [8] (cf. also [6]; Thron's theorem should be applied to the function $\hat{f}(x) = r - f(r-x)$ having the fixed point at zero).

THEOREM 3. *Let $\{a_n\}$ be a sequence of positive numbers fulfilling* (1) *and* (18). *If $pk = 1$, then for $n \geqslant 2$ the sequence $\{a_n\}$ is strictly decreasing and* (30) *holds true. Moreover,*

(34) $$a_n - r \sim c^{2^n}, \quad n \to \infty,$$

where $c \in (0, 1)$ is a constant depending on a_1.

Proof. Now we have

(35) $$z = u = r,$$

whence it follows in virtue of (11), (9) and (13) that $r < f(x) < x$ in (r, ∞). Thus $f \in S_r^0[(r, \infty)]$ and further we argue as in the proof of Theorem 2. Note that for $n \geqslant 2$ the sequence $\{a_n\}$, being strictly monotonic, has to be, in fact, strictly decreasing because of (15), (30), and (35). Condition (34) again is a consequence of the results in [8] (cf. also [6]).

Remark 2. When $pk = 1$, then algorithm (1) reduces to (3). Thus Theorem 3 can also be deduced from known properties of Newton's method (4) applied to $F(x) = x^k - N$; cf. [1], [2].

THEOREM 4. *Let $\{a_n\}$ be a sequence of positive numbers fulfilling* (1) *and* (18). *If $1 < pk \leqslant 2$ and f fulfils* (19) (cf. Lemma 3 and Remark 1), *then for large n the sequence $\{a_n\}$ oscillates around r and* (30) *holds true. Moreover,*

(36) $$|a_n - r| \sim |s|^n, \quad n \to \infty,$$

when $1 < pk < 2$, and

(37) $$|a_n - r| \sim 1/\sqrt{n}, \quad n \to \infty,$$

when $pk = 2$.

Proof. Now we have $u < r < z$. When x runs from z to infinity, $f(x)$ strictly increases from $u = f(z)$ to infinity. Consequently there exists a unique point $v > z$ such that $f(v) = r$. According to (16) we have

(38) $$v \in M.$$

In view of (15) we have $a_2 \geqslant u$. Suppose that $a_2 > v$. By (5) and (11), $a_3 = f(a_2) < a_2$. If $a_3 > v$ we can repeat this argument arriving thus after a finite number of steps at an a_m such that (cf. (18) and (38))

(39) $$a_m \in [u, v).$$

In fact, if we had $a_n > v$ for all $n \geqslant 2$ (by (18) and (38), $a_n \neq v$ for $n \in N$), then the sequence $\{a_n\}_{n \geqslant 2}$ would be strictly decreasing and thus it would converge to a limit $g \geqslant v > r$. Because of the continuity of f we would have (cf. [5]–[7]) $f(g) = g$, which is incompatible with (11). Thus there exists an $m \in N$ such that (39) is fulfilled.

We have by (10)

$$\lim_{x \to 0} f^2(x) = \infty, \quad \lim_{x \to \infty} x^{-1} f^2(x) = q^2 < 1,$$

which together with (19) implies that

(40) $$f^2(x) > x \quad \text{in } (0, r), \quad f^2(x) < x \quad \text{in } (r, \infty).$$

According to (13) the function f is strictly decreasing in $[u, z)$ from the value (cf. (40)) $f(u) = f^2(z) < z$ to the value $f(z) = u$, and is strictly increasing in (z, v) from $f(z) = u$ to $f(v) = r$. Consequently

(41) $$f(x) \in (r, v) \quad \text{for } x \in [u, r), \quad f(x) \in [u, r) \quad \text{for } x \in (r, v).$$

Hence

(42) $$f^2(x) < r \quad \text{in } [u, r), \quad f^2(x) > r \quad \text{in } (r, v).$$

Relations (40) and (42) show that $f^2 \in S_r^0[[u, v)]$, whence it follows in view of (39) and of Lemma 1 that for $n > m/2$ the sequences $\{a_{2n}\}$ and $\{a_{2n+1}\}$ are strictly monotonic and converge to r. Hence (30) results.

The oscillatory behaviour of $\{a_n\}$ for $n \geq m$ is a consequence of (39), (5) and (41).

When $1 < pk < 2$, then $-1 < s < 0$ and $(f^2)'(r) = s^2 \in (0, 1)$. We deduce from the results in [8] (cf. also [6]) that then

$$|a_{2n} - r| \sim s^{2n}, \quad |a_{2n+1} - r| \sim s^{2n}, \quad n \to \infty.$$

According to Definition 2 this means that there exist positive numbers g_1 and g_2 such that

$$\lim_{n\to\infty} s^{-2n}|a_{2n} - r| = g_1, \quad \lim_{n\to\infty} s^{-2n}|a_{2n+1} - r| = g_2. \tag{43}$$

Hence we obtain by (5) and (30)

$$\frac{g_2}{g_1} = \lim_{n\to\infty} \left|\frac{a_{2n+1} - r}{a_{2n} - r}\right| = \lim_{n\to\infty} \left|\frac{f(a_{2n}) - r}{a_{2n} - r}\right| = |s|$$

so that $g_2 = |s| g_1$. Thus we have by virtue of (43)

$$\lim_{n\to\infty} |s|^{-2n}|a_{2n} - r| = g_1 = \lim_{n\to\infty} |s|^{-(2n+1)}|a_{2n+1} - r|,$$

which yields (36).

When $pk = 2$ relation (37) can be derived in a similar manner.

As a byproduct we have the following

COROLLARY. *Under the assumptions of Theorem* 4 *the sequences* $\{a_{2n}\}$ *and* $\{a_{2n+1}\}$ *are strictly monotonic for large* n.

Remark 3. Relation (30) in Theorems 2, 3 and 4 may also be deduced from the contents of paper [3].

THEOREM 5. *Let* $\{a_n\}$ *be a sequence of positive numbers fulfilling* (1) *and* (18). *If* $pk > 2$, *then the sequence* $\{a_n\}$ *diverges. More exactly, it has neither finite nor infinite limit.*

Proof. We have $|s| > 1$, consequently the fixed point $x = r$ of f is repulsive (cf. [5], [7]) and (30) cannot be true. The sequence $\{a_n\}$ cannot converge to another point $g \in \mathbf{R}^+$ either for otherwise we would have $f(g) = g$ contrary to (11). By (15) the sequence $\{a_n\}$ cannot converge to zero. Suppose that $\lim_{n\to\infty} a_n = \infty$. Then there exists an $n_0 \in N$ such that

$$a_n > v, \quad n > n_0, \tag{44}$$

where v has the same meaning as in the proof of Theorem 4. But as we have seen, (44) leads to a contradiction.

Remark 4. Asymptotic relations (32), (34), (36) and (37) have a rather theoretical importance. They show, in particular, that the convergence (30) is fastest when $p = 1/k$. This is just the case where (1) coincides with Newton's algorithm (3). On the other hand, as may be seen from (37), when $p = 2/k$ the convergence can be very slow.

Remark 5. The strict monotonicity of the sequence $\{a_n\}$ for $n \geqslant 2$ asserted in Theorems 2 and 3 implies that for $n \geqslant 2$ every term of $\{a_n\}$ yields a better approximation of r than the previous one. However, this does not allow one to estimate the error of approximation $|a_n - r|$ at every step. Neither does the asymptotic relation (32) or (34), respectively. From this point of view we are in a much more favourable situation in cases covered by Theorem 4, where the convergence is ultimately oscillatory. Then it is still true that, for large n, each a_n yields a better approximation of r than does a_{n-1} (cf. the Corollary to Theorem 4), but due to the oscillatory character of the sequence $\{a_n\}$ we have the error estimate

$$|a_n - r| \leqslant |a_{n+1} - a_n|, \quad n > m, \tag{45}$$

where $m \in N$ is such that (39) is fulfilled. We will return to the problem of m in a while.

Thus if (19) is fulfilled the most convenient choice of p might be a value slightly larger than $1/k$ (at any rate $p < 2/k$). Then the convergence (30) is geometric (cf. (36); the closer p is to $1/k$, the smaller is $|s|$ and consequently the faster is the convergence (30)), and according to Theorem 4 the sequence $\{a_n\}$ is ultimately oscillatory so that we have the error estimate (45).

When we start the algorithm (1) with an a_1 chosen at random from R^+ we do not know which of the cases (17), (18) occurs. We proceed as if (18) were the case (which is by far much more probable). If we arrive at an $m \in N$ such that

$$a_{m+1} = a_m, \tag{46}$$

then we realize that in fact we have (17), the common value in (46) is r, and the stationary sequence $\{a_n\}$ (cf. Theorem 1) satisfies (29) (resulting from (31)) and hence also (45).

As we have seen in the proof of Theorem 4 (the argument remains essentially the same when (17) is fulfilled), if $1/k < p < 2/k$, then the sequence $\{a_n\}$ is strictly decreasing for $n = 2, \ldots, m$, where m is such that (39) is true. (If we have (17), then (39) and (46) are equivalent.) This suggests the following procedure: at every step of the algorithm (1) we calculate also the difference $a_{n+1} - a_n$. If for an index $m \geqslant 2$ we have

$$a_{m+1} - a_m \geqslant 0, \tag{47}$$

then (45) is valid.

Theoretically, if (19) is fulfilled and $1/k < p < 2/k$, then with a_1 chosen at random from R^+ the index m in (47) may be very large. The index m tends to infinity when a_1 approaches zero or infinity. But in practice we usually roughly know the approximate value of r and may choose a_1 reasonably close to r. Then also $m \geqslant 2$ fulfilling (47) will not be too large.

References

[1] S. Baranger, *Introduction à l'analyse numérique*, Hermann, Paris 1977.

[2] G. Dahlquist and Åke Björck, *Numerical Methods*, Prentice-Hall Series in Automatic Computation, Englewood Cliffs, N.J. 1974.

[3] J. Drewniak, J. Kalinowski and B. Ulewicz, *On the behaviour of continuous real functions in the neighbourhood of a fixed point*, Aequationes Math. 14 (1976), 123–136.

[4] P. Grogono, *Programming in Pascal*, Addison-Wesley, Reading, Mass. 1978.

[5] M. Kuczma, *Functional Equations in a Single Variable*, Monografie Mat. 46, Polish Scientific Publishers, Warszawa 1968.

[6] –, B. Choczewski and R. Ger, *Iterative Functional Equations*, Encyclopedia Math. Appl. 32, Cambridge University Press, Cambridge 1990.

[7] Gy. Targonoski, *Topics in Iteration Theory*, Vandenhoeck und Ruprecht, Göttingen 1981.

[8] W. J. Thron, *Sequences generated by iteration*, Trans. Amer. Math. Soc. 96 (1960), 38–53.

INSTYTUT MATEMATYKI, UNIWERSYTET ŚLĄSKI
DEPARTMENT OF MATHEMATICS, SILESIAN UNIVERSITY
Bankowa 14, 40-007 KATOWICE, POLAND

and

Kazimierza Wielkiego 3/42, 30-074 KRAKÓW, POLAND

Reçu par la Rédaction le 13.03.1989