INFINITE PRODUCT EXPANSIONS FOR MATRIX n-th ROOTS

R. A. SMITH

(Received 22 September 1966)

1. Introduction

In this paper a denotes a square matrix with real or complex elements (though the theorems and their proofs are valid in any Banach algebra). Its spectral radius $\rho(a)$ is given by

$$\rho(a) = \lim ||a^\nu||^{1/\nu}, \text{ as } \nu \to \infty, \tag{1}$$

with any matrix norm (see [4], p. 183). If $\rho(a) < 1$ and n is a positive integer then the binomial series

$$S(a) = \sum_{\nu=0}^{\infty} \binom{-1/n}{\nu} (-a)^\nu \tag{2}$$

converges and its sum satisfies $S(a)^n = (1-a)^{-1}$. Let

$$u(x) = 1+\sum_{\nu=1}^{q-1} \frac{\Gamma(n^{-1}+\nu)x^\nu}{\nu!\,\Gamma(n^{-1})}, \tag{3}$$

where q is any integer exceeding 1. Then $u(a)$ is the sum of the first q terms of the series (2). Write

$$f(x) = 1+u(x)^n(x-1) \tag{4}$$

and let $a_0, a_1, a_2, \cdots$ be the sequence of matrices obtained by the iterative procedure

$$a_0 = a, \quad a_{\nu+1} = f(a_\nu). \tag{5}$$

Defining polynomials $\phi_0(x), \phi_1(x), \phi_2(x), \cdots$ inductively by

$$\phi_0(x) = x, \quad \phi_{\nu+1}(x) = f(\phi_\nu(x)), \tag{6}$$

we have $a_\nu = \phi_\nu(a)$ and therefore $a_\mu a_\nu = a_\nu a_\mu$ for all μ, ν. The following is proved in section 2:

THEOREM 1. *If* $\rho(a) < 1$ *then*

$$P(a) = \prod_{\nu=0}^{\infty} u(a_\nu) \tag{7}$$

converges and $P(a) = S(a)$. *Furthermore, if* $\rho(a) < r < 1$, *then*

$$\|a_\nu\| < Mr^{q^\nu} \tag{8}$$

for all ν, *where* M *depends on* r *and* a *but is independent of* ν *and* q.

Inequality (8) shows that $P(a)$ converges very rapidly. This could make it useful for the numerical computation of $S(a)$. In general the series (2) converges too slowly to be used for this purpose. In section 3 it is shown that when $n > 1$ the infinite product (7) converges for a larger class of matrices a than does the series (2). If $n = 1$ then (3) and (4) give $f(x) = x^q$. The solution of (5) is then $a_\nu = a^{q^\nu}$ and (7) reduces to

$$(1-a)^{-1} = \prod_{\nu=0}^{\infty} \{1+a^{q^\nu}+a^{2q^\nu}+\cdots+a^{(q-1)q^\nu}\}. \tag{9}$$

This well-known formula goes back to Euler. Its use for practical computation was suggested by Ostrowski [6], Hotelling [3] and others. Hotelling was able to connect (9) in the special case $q = 2$ with an iterative method for matrix inversion given by the Newton-Raphson formula. For (7) there is a similar connection with the Newton-Raphson formula which is discussed in section 4. Theorem 1 can be used to find a matrix c satisfying $c^n = b$ for any square matrix b whose spectrum lies entirely in the half plane $\operatorname{Re}\lambda > 0$. For the spectrum of $a = (b+1)^{-1}(b-1)$ then lies in the disc $|\lambda| < 1$ and $c = (1+a)^{1/n}(1-a)^{-1/n}$ can be computed with the help of (7). In the special case when the eigenvalues of b are real and positive it is simpler to take $a = 1-k^{-1}b$, where k is any real number satisfying $k > \frac{1}{2}\rho(b)$. The eigenvalues of a then satisfy $-1 < \lambda < 1$ and $c = k^{1/n}(1-a)^{1/n}$ can be computed with the help of (7).

2

LEMMA 1. *If* m *is any integer in the range* $1 \leqq m \leqq n$ *and*

$$u(x)^m = \beta_0+\beta_1 x+\beta_2 x^2+\cdots$$

then for all ν,

$$\beta_\nu \geqq \beta_{\nu+1} \geqq 0. \tag{10}$$

PROOF. Since $(1-x)^{-1/n} = u(x)+x^q v_1(x)$, it follows that

$$(1-x)^{-m/n} = u(x)^m+x^q v_m(x), \tag{11}$$

where $v_1(x)$, $v_m(x)$ are power series with positive coefficients. Comparing coefficients in (11) we get $(-1)^\nu\binom{-m/n}{\nu} = \beta_\nu$ for $0 \leqq \nu < q$. If $1 \leqq m \leqq n$ then $(-1)^\nu\binom{-m/n}{\nu}$ is a positve monotonic decreasing function of ν. Hence, (10) holds in the range $0 \leqq \nu < q-1$ and its remains to prove (10) for the range $\nu \geqq q-1$. We do this by induction over m. Clearly (10) holds when $m = 1$ because then $\beta_\nu = 0$ for all $\nu \geqq q$. If the lemma holds for some m in the range $1 \leqq m < n$ then (3) gives

$$u(x)^{m+1} = u(x)^m u(x) = \Big(\sum_{\nu=0}^{\infty} \beta_\nu x^\nu\Big)\Big(\sum_{\nu=0}^{q-1} \alpha_\nu x^\nu\Big) = \sum_{\nu=0}^{\infty} \gamma_\nu x^\nu,$$

where

$$\alpha_\mu = \frac{\Gamma(n^{-1}+\mu)}{\mu!\,\Gamma(n^{-1})} \text{ and } \gamma_\nu = \sum_{\mu=0}^{q-1} \alpha_\mu \beta_{\nu-\mu}$$

for all $\nu \geqq q-1$. Since (10) holds for all ν we have

$$0 \leqq \gamma_{\nu+1} = \sum_{\mu=0}^{q-1} \alpha_\mu \beta_{\nu+1-\mu} \leqq \sum_{\mu=0}^{q-1} \alpha_\mu \beta_{\nu-\mu} = \gamma_\nu,$$

for all $\nu \geqq q-1$. The lemma is therefore true for $m+1$ also. This establishes Lemma 1.

Lemma 2. *$f(x) = x^q g(x)$ and $\phi_\nu(x) = x^{q^\nu}\psi_\nu(x)$ where $g(x)$, $\psi_\nu(x)$ are polynomials with real non-negative coefficients which satisfy $g(1) = \psi_\nu(1) = 1$.*

Proof. With $m = n$, (11) gives

$$x^q v_n(x)(1-x) = 1+u(x)^n(x-1) = f(x).$$

Hence $f(x) = x^q g(x)$ for some polynomial $g(x)$. If $u(x)^n = \beta_0+\beta_1 x+\beta_2 x^2+\cdots$ then $\beta_0 = 1$ and $x^q g(x) = 1+u(x)^n(x-1) = \sum_{\nu=0}^{\infty} (\beta_\nu-\beta_{\nu+1})x^\nu$.

The coefficients of $g(x)$ are therefore non-negative by Lemma 1. Also $g(1) = f(1) = 1$ from (4). Induction over ν will be used to prove that $\phi_\nu(x)$ is of the form $x^{q^\nu}\,\psi_\nu(x)$, where $\psi_\nu(x)$ has non-negative coefficients. This is trivial for $\nu = 0$ since $\phi_0(x) = x$. If it is true for some integer ν then (6) gives

$$\phi_{\nu+1} = (\phi_\nu)^q g(\phi_\nu) = (x^{q^\nu}\psi_\nu)^q g(\phi_\nu) = x^{q^{\nu+1}}\psi_{\nu+1},$$

where $\psi_{\nu+1} = (\psi_\nu)^q g(\phi_\nu)$ is a polynomial with non-negative coefficients. The result is therefore true for $\nu+1$ also and the induction is complete. Since $f(1) = 1$ it follows from (6) by induction that $\phi_\nu(1) = 1$ for all ν. Hence $\psi_\nu(1) = \phi_\nu(1) = 1$ and the proof of Lemma 2 is finished.

When $q = 2$, (3) gives $u(x) = 1+n^{-1}x$. Then (4) gives

$$\begin{aligned} f &= 1+nu^{n+1}-(n+1)u^n, \\ &= (u-1)^2(1+2u+3u^2+\cdots+nu^{n-1}), \\ &= x^2 n^{-2}(1+2u+3u^2+\cdots+nu^{n-1}). \end{aligned} \tag{12}$$

This is the relation $f(x) = x^q g(x)$ for the special case $q = 2$. If x is small then $f(x)$ can be computed more accurately from the relation $f(x) = x^q g(x)$ than from (4) which involves internal cancellation when x is small.

PROOF OF THEOREM 1. If $\rho(a) < r$ then (1) gives

$$\|a^\nu\| \leqq Mr^\nu \tag{13}$$

for some constant M. Hence $\|\phi_\nu(a)\| \leqq M\phi_\nu(r)$ since the coefficients of $\phi_\nu(x)$ are non-negative by Lemma 2. Since $0 < r < 1$, Lemma 2 also gives $\|\phi_\nu(a)\| \leqq Mr^{q^\nu}\psi_\nu(r) \leqq Mr^{q^\nu}\psi_\nu(1) = Mr^{q^\nu}$.

This proves (8) because $a_\nu = \phi_\nu(a)$. From (4) and (5),

$$(1-a_{\nu+1}) = u(a_\nu)^n(1-a_\nu).$$

Since $a_\mu a_\nu = a_\nu a_\mu$ it follows by induction that

$$(1-a_{\kappa+1}) = [\prod_{\nu=0}^{\kappa} u(a_\nu)]^n(1-a). \tag{14}$$

Since $a_\nu = \phi_\nu(a)$ we have $\prod_{\nu=0}^{\kappa} u(a_\nu) = W_\kappa(a)$, where $W_\kappa(x)$ is a polynomial with non-negative coefficients. When $-1 < x < 1$ it follows from (2) and (14) that

$$S(x) = (1-x)^{-1/n} = W_\kappa(x)\{1-\phi_{\kappa+1}(x)\}^{-1/n}. \tag{15}$$

Expanding $\{1-\phi_{\kappa+1}(x)\}^{-1/n}$ by the binomial theorem and using Lemma 2 we get

$$S(x) = W_\kappa(x)+x^{q^{\kappa+1}}V_\kappa(x), \tag{16}$$

where $V_\kappa(x)$ is a power series with non-negative coefficients. This and (13) give $\|S(a)-W_\kappa(a)\| = \|a^{q^{\kappa+1}}V_\kappa(a)\| \leqq Mr^{q^{\kappa+1}}V_\kappa(r)$. Since none of the coefficients of $S(x)$ exceeds 1 by (2), the same is true of the coefficients of $V_\kappa(x)$ by (16). Hence,

$$\begin{gathered} \|S(a)-W_\kappa(a)\| \leqq Mr^{q^{\kappa+1}}(1-r)^{-1}, \\ S(a) = \lim_{\kappa\to\infty} W_\kappa(a) = \prod_{\nu=0}^{\infty} u(a_\nu). \end{gathered} \tag{17}$$

This completes the proof of Theorem 1.

3. Domain of convergence

Let $D = \bigcup_{\nu=0}^{\infty} D_\nu$ where D_ν is the set of points in the complex z plane for which $|\phi_\nu(z)| < 1$. Each D_ν is an open set and D_0 is the disc $|z| < 1$.

THEOREM 2. *If the spectrum of a lies wholly in D then the infinite product* (7) *converges and satisfies* $P(a)^n = (1-a)^{-1}$.

PROOF. Lemma 2 gives $|f(z)| < g(|z|) < g(1) = 1$ for $|z| < 1$.

This and (6) show that $|\phi_{\nu+1}(z)| < 1$ when $|\phi_\nu(z)| < 1$. That is,

$D_{\nu+1} \supset D_\nu$ for all ν. Since the spectrum of a is compact it must lie in D_μ for some μ. Then the spectrum of $a_\mu = \phi_\mu(a)$ lies wholly in the disc $|z| < 1$. Hence $\rho(a_\mu) < 1$ and $S(a_\mu) = \prod_{\nu=\mu}^{\infty} u(a_\nu)$ by Theorem 1. From (14) with $\kappa = \mu - 1$ we get

$$P(a)^n = S(a_\mu)^n \left[\prod_{\nu=0}^{\mu-1} u(a_\nu)\right]^n = S(a_\mu)^n(1-a_\mu)(1-a)^{-1} = (1-a)^{-1}.$$

This completes the proof of Theorem 2. If $n = 1$ then $\phi_\nu(z) = z^{q^\nu}$ and $D_\nu = D_0$ for all ν. That this is not so when $n > 1$ is shown by the next theorem.

THEOREM 3. *If $n > 1$ then D_1 includes all of the closed disc $|z| \leqq 1$ except the point $z = 1$.*

PROOF. Lemma 2 gives $|f(z)| \leqq f(|z|) \leqq f(1) = 1$ for $|z| \leqq 1$. The inequality $|f(z)| \leqq f(|z|)$ can reduce to equality only when the terms of $f(z)$ all have the same complex argument (see [2], p. 26). If

$$u(z) = k_1 z^{q-1} + k_2 z^{q-2} + \cdots$$

then (4) gives

$$f(z) = k_1^n z^{1+n(q-1)} + (nk_2 - k_1)k_1^{n-1} z^{n(q-1)} + \cdots,$$

where the terms shown are those of the highest degrees. If $n > 1$ then both these terms have positive coefficients because $k_2 \geqq k_1 > 0$ by (3). These terms have the same complex argument only when z is real and positive. Therefore $|f(z)| \leqq f(|z|)$ reduces to equality only when z is real and positive. Hence $z = 1$ is the only point of the disc $|z| \leqq 1$ at which $|f(z)| \leqq 1$ reduces to equality. Since $f(z) = \phi_1(z)$ it follows that D_1 includes all of the disc $|z| \leqq 1$ except the point $z = 1$. This established Theorem 3.

Since D_1 is an open set Theorem 3 shows that a part of it must lie outside the circle $|z| = 1$ when $n > 1$. The region of convergence of the infinite product $P(a)$ is therefore somewhat larger than that of the series (2) which diverges if a has an eigenvalue outside the circle $|z| = 1$. More precise information about the size of D will be given only for the special case $q = 2$. Let H be the convex hull of the set which is the union of the closed disc $|z| \leqq 1$ and the single point $z = -n$. Let H_0 be the set obtained by deleting from H the two points $z = -n, 1$.

THEOREM 4. *If $q = 2$ and $n > 1$ then D_1 includes H_0.*

PROOF. If $z + n = de^{i\delta}$ then $0 < d < n+1$ and

$$-\frac{1}{n} \leqq \sin\delta \leqq \frac{1}{n} \tag{18}$$

for all z in H_0. Also $u(z) = 1 + n^{-1}z = n^{-1}de^{i\delta}$ since $q = 2$. From (12) we get

https://doi.org/10.1017/S1446788700005309 Published online by Cambridge University Press

$$|f(z)|^2 = \{1+nu^{n+1}-(n+1)u^n\}\{1+n\bar{u}^{n+1}-(n+1)\bar{u}^n\}.$$

With $u = n^{-1}de^{i\delta}$ this gives $|f(z)|^2 = 1+(d/n)^n h(d, \delta)$ where

$$\begin{aligned} h(d, \delta) = {} & (d/n)^n\{d^2-2(n+1)d\cos\delta+(n+1)^2\} \\ & +2d\cos(n+1)\delta-2(n+1)\cos n\delta. \end{aligned}$$

Hence $|f(z)| < 1$ if and only if $h(d, \delta) < 0$. Since $f(z) = \phi_1(z)$ it follows that $z \in D_1$ if and only if $h(d, \delta) < 0$. To prove Theorem 4 it is therefore sufficient to show that $h(d, \delta) < 0$ throughout H_0. Clearly $h(d, \delta) < 0$ in $H_0 \cap \bar{D}_0$ since D_1 includes the closed disc $\bar{D}_0$, except for the point $z = 1$, by Theorem 3. To prove that $h(d, \delta) < 0$ in the whole of H_0 it is therefore sufficient to show that $\partial h/\partial d > 0$ in $H_0-\bar{D}_0$ because each point of $H_0-\bar{D}_0$ lies on some line segment joining $z = -n$ to a point of $H_0 \cap \bar{D}_0$. Since $|z|^2 = |-n+de^{i\delta}|^2 = n^2-2nd\cos\delta+d^2$, we can express $\partial h/\partial d$ in the form

$$\partial h/\partial d = n^{-(n+1)}d^{n-1}\{(n+1)^2|z|^2-d^2\}+2\cos(n+1)\delta. \tag{19}$$

Since $x^{-1}\sin x \geqq 3/\pi$ in $0 < x \leqq \pi/6$, (18) gives

$$\sin|\delta| \leqq n^{-1} \leqq 3(2n+2)^{-1} \leqq \sin\{(2n+2)^{-1}\pi\}.$$

Hence $|\delta| \leqq (2n+2)^{-1}\pi$ and $\cos(n+1)\delta \geqq 0$ for all z in H_0. This and (19) give

$$\partial h/\partial d \geqq n^{-(n+1)}d^{n-1}\{(n+1)^2|z|^2-d^2\} > n^{-(n+1)}d^{n-1}(n+1)^2\{|z|^2-1\},$$

for all z in H_0. Therefore $\partial h/\partial d > 0$ in $H_0-\bar{D}_0$. This completes the proof of Theorem 4. Notice that the points $z = -n$, 1 which were omitted from H_0 lie outside D because $f(-n) = 1$ when $q = 2$ and therefore $\phi_\nu(-n) = \phi_\nu(1) = 1$ for all $\nu \geqq 1$ by (6).

4. Connection with Newton-Raphson

The Newton-Raphson formula for the numerical solution of an equation $Y(x) = 0$ is $x_\nu - x_{\nu+1} = Y(x_\nu)/Y'(x_\nu)$. With $Y(x) = b-x^{-n}$ this becomes

$$x_{\nu+1} = x_\nu\{1+n^{-1}(1-bx_\nu^n)\}. \tag{20}$$

As a generalisation of this we consider the formula

$$x_{\nu+1} = x_\nu u(a_\nu), \quad a_\nu = 1-bx_\nu^n, \tag{21}$$

where $u(x)$ is given by (3) with any q. This reduces to (20) when $q = 2$ because then $u(x) = 1+n^{-1}x$. When x_ν and b are square matrices, (21) gives

$$x_{\nu+1} = x_0 u(a_0)u(a_1)\cdots u(a_\nu). \tag{22}$$

Also, (4) and (21) show that $f(a_\nu) = 1+(a_\nu-1)u(a_\nu)^n = 1-bx_\nu^n u(a_\nu)^n$.

If $x_\nu a_\nu = a_\nu x_\nu$ then $x_\nu^n u(a_\nu)^n = x_{\nu+1}^n$ by (21) and

$$f(a_\nu) = 1-b(x_{\nu+1})^n = a_{\nu+1}. \tag{23}$$

Compare this with (5). The condition $x_\nu a_\nu = a_\nu x_\nu$ is satisfied if $x_\nu b = bx_\nu$ and this is true by induction provided that $x_0 b = bx_0$. When this is so, (22), (23) and Theorem 1 show that $x_{\nu+1} \to x_0 P(a_0)$ as $\nu \to \infty$ provided that $\rho(a_0) < 1$. If $x_0 P(a_0) = L$ then

$$L^n = x_0^n P(a_0)^n = x_0^n (1-a_0)^{-1} = b^{-1}.$$

The following theorem is therefore true.

THEOREM 5. *If $a_0 = 1-bx_0^n$ has $\rho(a_0) < 1$ and $x_0 b = bx_0$ then the sequence of matrices $x_0, x_1, x_2, \cdots$ obtained from (21) tends to a limit matrix L which satisfies $L^n = b^{-1}$. Furthermore, the rate of convergence is of the q-th order.*

Altman [1] and Petryshyn [7] have studied (21) in the special case when $n = 1$. They obtain results similar to Theorem 5 but without the requirement $x_0 b = bx_0$. This requirement can be deleted from Theorem 5 in the case $n = 1$ because (23) then follows without use of the relation $x_\nu a_\nu = a_\nu x_\nu$. The following counter-example shows that $x_0 b = bx_0$ cannot be deleted from Theorem 5 when $n > 1$. If

$$b = \begin{pmatrix} 1 & 0 \\ 0 & \mu^{-n} \end{pmatrix}, \quad x_\nu = \begin{pmatrix} 1 & \xi_\nu \\ 0 & \mu \end{pmatrix}, \quad x_\nu^n = \begin{pmatrix} 1 & \zeta\xi_\nu \\ 0 & \mu^n \end{pmatrix}$$

then $\zeta = (\mu-1)^{-1}(\mu^n-1)$ and $\rho(a_0) = 0$ where $a_0 = 1-bx_0^n$. These matrices satisfy (20) provided that $\xi_\nu = (1-n^{-1}\zeta)^\nu \xi_0$ for all ν. If $n > 1$ and $\mu \geqq 4$ then $|1-n^{-1}\zeta| > 1$ and x_ν does not tend to a limit as $\nu \to \infty$ because $|\xi_\nu| \to \infty$. The deletion of $x_0 b = bx_0$ from Theorem 5 therefore produces a false proposition when $n > 1$. When $n > 1$ and $q = 2$, Theorem 4 enables the condition $\rho(a_0) < 1$ in Theorem 5 to be replaced by the requirement that the spectrum of a_0 lie wholly in H_0.

With $Y(x) = bx^n-1$ the Newton-Raphson formula becomes

$$x_{\nu+1} = (1-n^{-1})x_\nu + (nbx_\nu^{n-1})^{-1}. \tag{24}$$

A higher order formula of Traub [8] generalises this is the same way that (21) generalises (20). Laasonen [5] has shown that (24) can be used to find a square root of any real matrix whose eigenvalues are all positive. Each iteration of (24) requires a matrix inversion which could introduce considerable error when $b^{-1/n}$ is ill-conditioned. The iterations of (5) and (21) do not involve matrix inversions.

References

[1] M. Altman, 'An optimum cubically convergent iterative method for inverting a linear bounded operator in Hilbert space', *Pacific J. Math.* 10 (1960), 1107–1113.
[2] G. H. Hardy, J. E. Littlewood and G. Polya, *Inequalities* (C.U.P., 2nd ed. 1952).
[3] H. Hotelling, 'Some new methods in matrix calculation', *Ann. Math. Statist.* 14 (1943), 1–34.
[4] A. S. Householder, *The theory of matrices in numerical analysis* (Blaisdell, New York, 1964).
[5] P. Laasonen, 'On the iterative solution of the matrix equation $AX^2-I = O$', *Math. Tables Aids Comput.* 12 (1958), 109–116.
[6] A. Ostrowski, 'Sur quelques transformations de la série de Liouville-Neumann', *C.R. Acad. Sci. Paris* 206 (1938), 1345–1347.
[7] W. V. Petryshyn, 'On the inversion of matrices and linear operators', *Proc. Amer. Math. Soc.* 16 (1965), 893–901.
[8] J. F. Traub, 'Comparison of iterative methods for the calculation of n-th roots', *Comm. ACM* 4 (1961), 143–145.

Department of Mathematics
University of Durham, England