

Solution of Equations Involving Analytic Functions

M. P. CARPENTIER AND A. F. DOS SANTOS

*Centro de Análise e Processamento de Sinais,
Instituto Superior Técnico, 1096 Lisboa Codex, Portugal*

Received March 4, 1981; revised September 2, 1981

The location, by the method of Delves and Lyness, of the roots of the equation $f(z) = 0$ for analytic f is examined. New formulas are proposed for calculating the integrals involved in the application of the method. These formulas do not require the evaluation of $f'(z)$ and are more accurate than some existing ones. A highly reliable, but not completely foolproof, procedure for computing the number of zeros of f in a given region is also proposed.

1. INTRODUCTION

In the fields of physics and applied mathematics one, very often, encounters the problem of determining all the roots of an equation of the form $f(z) = 0$ in some region of the z plane; examples of this are the study of the characteristic modes of waveguides and the calculation of characteristic frequencies of resonating systems.

The most commonly used routines for finding the roots of the equation $f(z) = 0$ (f analytic) do not ensure that all the zeros of f in the region of interest are obtained, except when f is a polynomial. To the authors' knowledge two methods, Gardiol's [1] and Delves and Lyness [2], have been proposed for the solution of this problem; the method of Lehmer [3] for solving polynomial equations can also be used for equations involving other analytic functions but we feel that the computation time needed will, in the latter case, be prohibitive. Recently a graphical technique [4] has also been suggested for the same purpose but it cannot compete in accuracy with any of the above mentioned methods.

Gardiol's method is based on the use of sequences derived from Newton's method starting from appropriate values on the boundary of the region. It is easy to use but has the serious drawback of not ensuring that all the zeros of f are found. The method proposed by Delves and Lyness is based on an entirely different principle which consists in determining a polynomial having the same zeros as f in the given region, the coefficients of the polynomial being calculated from the integrals of $z^k f'(z)/f(z)$ along the boundary of the region for a number of values of k equal to the number of zeros of f . Providing that the number of zeros of f is correctly calculated the method is reliable and is equally applicable to simple and multiple zeros. This method has been applied to the solution of physical problems such as the calculation of eigenvalues corresponding to the modes supported by inhomogeneous waveguides [5, 6]

and the study of plasma instabilities [7], but it does not seem to be as widely used as it would be expected bearing in mind its advantages.

In the present paper we give new formulas for calculating the above mentioned integrals. We also propose a test which significantly reduces the possibility for an error to occur in the calculation of the number of zeros of f , a question of fundamental importance in the application of the method. The formulas given do not require the calculation of the derivative of f which is often impracticable or may involve an amount of computation substantially higher than that of f . The application of these formulas is restricted to circular integration paths but this is not thought to be a serious limitation.

2. THE ASSOCIATED POLYNOMIAL

Let f be a function with n zeros z_j ($j = 1, 2, \dots, n$) in a bounded simply-connected region X of the complex plane and analytic in $X + \Gamma$, where Γ is the boundary of X (some of z_j may be equal). Let

$$P(z) = \sum_{k=0}^n a_k z^k \quad (a_n = 1) \quad (1)$$

be the polynomial of degree n whose zeros are equal to and have the same multiplicity as the zeros of f in X . Henceforth this polynomial will be referred to as the associated polynomial for the region X . Multiplying (1) by $z^{-m}P'(z)/P(z)$ and integrating along a path Γ_1 enclosing the origin and all the zeros of $P(z)$, the following set of equations is obtained:

$$ma_m = \sum_{k=0}^n a_k S_{k-m} \quad (m = 0, 1, \dots, n-1), \quad (2)$$

where

$$S_k = \frac{1}{2\pi i} \oint_{\Gamma_1} \frac{P'(z)}{P(z)} z^k dz. \quad (3)$$

Formulas (2) which constitute the basis of the method of Delves and Lyness are known in the literature as Newton's formulas and can be derived in other ways [8]. For negative m the above procedure leads to

$$0 = \sum_{k=0}^n a_k S_{k-m}$$

which we shall not use in the following.

In the integrals (3) the integration path encloses all poles of $z^k P'(z)/P(z)$ and thus

$$L_k = \frac{1}{2\pi i} \oint \frac{f'(z)}{f(z)} z^k dz \quad (6)$$

calculated along the boundaries of X_{m+1} and X_m . Henceforth the regions X_m are chosen to be circles centred at the origin. Denoting by ρ_m the radius of X_m we have

$$\rho_m^{-k} L_k = C_k = \frac{1}{2\pi} \int_0^{2\pi} \frac{-i\phi'(\theta)}{\phi(\theta)} e^{ik\theta} d\theta, \quad (7)$$

where $\phi(\theta) = f(\rho_m e^{i\theta})$ and C_k is the Fourier coefficient of order k of the function $-i\phi'(\theta)/\phi(\theta)$, i.e.,

$$-i \frac{\phi'(\theta)}{\phi(\theta)} = \sum_r C_r e^{-ir\theta}. \quad (8)$$

Formulas (6) and (7) show that the Fourier coefficient C_0 represents the number of zeros of f in X_m . For any C_k the application of the trapezoidal rule with $N + 1$ points yields the formula

$$C_k^{(N)} = \frac{1}{N} \sum_{l=1}^N \frac{-i\phi'(\theta_l)}{\phi(\theta_l)} e^{ik\theta_l}, \quad (9)$$

where $\theta_l = (2\pi/N)l$. The approximation error is

$$\varepsilon_k^{(N)} = C_k^{(N)} - C_k = \sum_q' C_{k+qN}, \quad (10)$$

where the symbol $'$ indicates that the term $q = 0$ is excluded from the summation.

However, the fact that the primitive of $\phi'(\theta)/\phi(\theta)$ is known permits the derivation of a formula for C_k that does not require knowledge of the derivative of f which is often impracticable to calculate. To this end we integrate (8) along the interval $[\theta - 2\pi/N, \theta]$ obtaining

$$\ln \left[\phi(\theta)/\phi \left(\theta - \frac{2\pi}{N} \right) \right] = \sum_r' \frac{C_r}{r} (e^{ir2\pi/N} - 1) e^{-ir\theta} + 2\pi i C_0/N. \quad (11)$$

The function in the LHS of (11) is the branch of $\ln[f(z)/f(ze^{-i2\pi/N})]$ that is analytic in an annular region containing the circumference Γ_m and whose imaginary part tends to zero when $N \rightarrow \infty$. The analyticity of the logarithm on the contour Γ_m is due to the zeros of $f(z)$ and $f(ze^{-i2\pi/N})$ in the region X_m being in one to one correspondence.

From (11) it follows that

$$C_0 = \frac{1}{2\pi i} \cdot \frac{N}{2\pi} \int_0^{2\pi} g(\theta) d\theta,$$

$$C_k = \frac{k}{\exp(ik(2\pi/N)) - 1} \cdot \frac{1}{2\pi} \int_0^{2\pi} g(\theta) e^{ik\theta} d\theta \quad (k \neq 0),$$

where

$$g(\theta) = \ln \left[\phi(\theta)/\phi \left(\theta - \frac{2\pi}{N} \right) \right].$$

The evaluation of these integrals by the trapezoidal rule yields the expressions

$$C_0 = \frac{1}{2\pi i} \sum_{l=1}^N g(\theta_l) = \frac{1}{2\pi} \sum_{l=1}^N \text{Im}[g(\theta_l)], \quad (12)$$

$$\tilde{C}_k^{(N)} = \frac{k/N}{\exp(ik(2\pi/N)) - 1} \sum_{l=1}^N e^{ik\theta_l} g(\theta_l) \quad (k \neq 0). \quad (13)$$

Note that formula (12) is exact. However, as is shown in Section 4, some difficulties arise in the computation of C_0 since knowledge of $g(\theta)$ at the points θ_l ($l = 1, 2, \dots, N$) may be insufficient to ensure that the computed values of $\ln [g(\theta)]$ belong to the same branch of this function.

The approximation error of (13) is easily obtained by putting $\theta = \theta_l$ in (11), multiplying both sides by $\exp(ik\theta_l)$ and summing over l . Denoting this error by $\delta_k^{(N)}$ we have:

$$\delta_k^{(N)} = \tilde{C}_k^{(N)} - C_k = \sum_q' k \frac{C_{k+qN}}{k + qN}. \quad (14)$$

Comparison of (10) and (14) shows clearly that, providing $k \ll N$, $|\delta_k^{(N)}|$ is less than $|\varepsilon_k^{(N)}|$, i.e., expression (13) is more accurate than (9). This is illustrated in Table I where the computed values of $C_k^{(N)}$ and $\tilde{C}_k^{(N)}$ are given for the function $f(z) = \sin(\pi z - \pi/4)$.

TABLE I
Computed Values of $C_k^{(N)}$ and $\tilde{C}_k^{(N)}$ for $f(z) = \sin(\pi z - \pi/4)$

ρ_m	L_0	$N/2^a$	$C_1^{(N)}$	$\tilde{C}_1^{(N)}$	$C_{10}^{(N)}$	$\tilde{C}_{10}^{(N)}$
0.8	2	32	-0.640317	-0.625234	0.533038	0.525619
		64	-0.625242	-0.625002	0.524605	0.524479
		≥ 128	-0.625000	-0.625000	0.524469	0.524469
1.0	2	16	-0.501066	-0.499970	0.0489311	0.0596729
		32	-0.500001	-0.500000	0.0563086	0.0563156
		≥ 64	-0.500000	-0.500000	0.0563145	0.0563145
10.0	20	64	-0.584950	-0.499953	1.77223	1.80442
		128	-0.503342	-0.499999	1.79634	1.79760
		256	-0.500006	-0.500000	1.79746	1.79746
		≥ 512	-0.500000	-0.500000	1.79746	1.79746

^aSince $f(z^*) = f^*(z)$ the integration can be reduced to the interval $[0, \pi]$. $N/2$ is the corresponding number of points.

At this point it is appropriate to point out that a formula of the type (13), i.e., not requiring the calculation of f' , is given by Delves and Lyness [2]. However, (13) is formally simpler and should be more accurate in most cases. In fact, g is a periodic infinitely differentiable function of θ which, as is known, leads to the best accuracy for the trapezoidal rule. The function considered by Delves and Lyness is only required to be continuous.

To end this section we note that, if a given accuracy is to be imposed on C_k , a straightforward application to formula (13) of the usual technique of doubling N until the specified accuracy is attained, is inconvenient in view of the fact that the coefficients $g(\theta_l)$ in (13) depend on N . The difficulty is overcome in the following way. Let M be the number of subintervals used in the computation of the index and Q the number of subdivisions of each subinterval at some stage in the computation (Q is a power of 2). Denoting by $E_k^{(N)}$ the summation in (13), we have

$$\begin{aligned} E_k^{(N)} &= \sum_{l=1}^N e^{ik(2\pi/N)l} \ln[\phi(2\pi l/N)/\phi(2\pi(l-1)/N)] \\ &= S_k^{(M,Q)} - e^{ik2\pi/N} (S_k^{(M,Q)} - E_k^{(M)}), \end{aligned}$$

where

$$\begin{aligned} S_k^{(M,Q)} &= \sum_{m=0}^{M-1} e^{ik(2\pi/M)m} \sum_{q=1}^Q e^{ik(2\pi/N)q} \\ &\quad \times \ln \left[\phi \left(\frac{2\pi}{N} q + \frac{2\pi}{M} m \right) / \phi \left(\frac{2\pi}{M} m \right) \right]. \end{aligned}$$

In the expression of $S_k^{(M,Q)}$ the inner summation can be performed following the usual trapezoidal rule algorithm. The calculation process starts with $N = M$, i.e., $Q = 1$ for which $S_k^{(M,1)} = E_k^{(M)}$.

4. COMPUTATION OF THE INDEX OF f

Let us now consider the evaluation on the computer of the number of zeros of f in a given region X (index of f in X). As pointed out in the preceding section, formula (12) is exact, providing $g(\theta)$ is the branch of $\ln[f(z)/f(ze^{-i2\pi/N})]$ that is analytic in an annulus containing the contour and vanishes for $N = \infty$. However, the range of $\text{Im}[g(\theta)]$ may be larger than $[-\pi, \pi]$ in which case the computer may give erroneous values for some points θ_l as it can only obtain the principal value of $\arg[f(z)/f(ze^{-i2\pi/N})]$. By increasing the number of points N , the length of the interval spanned by $\text{Im}[g(\theta)]$ decreases and hence for N greater than some N_0 the computed L_0 is correct.

Bearing in mind the principle of the argument it is easy to see that, for $\arg[\phi(\theta_l)/\phi(\theta_{l-1})]$ to lie in the interval $[-\pi, \pi]$ for all l , it is necessary that $N \geq 2L_0$, where equality corresponds to a uniform variation of $\arg[\phi(\theta)/\phi(\theta - 2\pi/N)]$ along the

contour. However, if a zero is close to the contour we may need $N \gg 2L_0$ to obtain a correct result.

Of course the above difficulties disappear if the index is computed through the formula

$$L_0 = \frac{1}{2\pi i} \oint_r \frac{f'(z)}{f(z)} dz$$

since the integrand is, in this case, a single-valued expression. However, apart from the inconvenience resulting from having to calculate the derivative of f , the trapezoidal rule algorithm is slowly convergent if a zero of f lies close to the contour and errors may occur if the condition for termination of the sequence of computed values of L_0 is not sufficiently restrictive.

Algorithms for calculating the index of analytic functions exist which make use of the fact that $\arg[f(z)]$ varies continuously along the contour (see [9, 10]) but none of these is safe from errors resulting from the presence of zeros close to the contour. A foolproof test to prevent the occurrence of these errors does not seem possible to devise. In the following we propose a simple test based on a criterion of proximity of the zero to the contour which may only fail in very anomalous cases.

Let α be the argument of $f(z)/f(ze^{-i2\pi/N})$ corresponding to the above specified branch of $\ln[g(\theta)]$ and $\tilde{\alpha}_l$ the value obtained on the computer for $\theta = \theta_l$. As pointed out above, an erroneous value of the index may be obtained if $|\alpha|$ exceeds π on any evaluation point; however, if the condition

$$|\tilde{\alpha}_l| < \alpha_0 < \pi \quad (l = 1, 2, \dots, N) \quad (15)$$

is imposed for some specified α_0 , an error can only occur if

$$|\alpha| > 2\pi - \alpha_0 \quad (16)$$

at some evaluation point. In fact if $\pi < |\alpha_l| \leq 2\pi - \alpha_0$ for some l then $\alpha_l \neq \tilde{\alpha}_l$ but this value is rejected by the computer since condition (15) is violated.

For example, if $f(z) = z - z_1$ the result is necessarily correct if we choose $\alpha_0 \leq \pi - \pi/N$, as can be seen from a simple geometrical reasoning. But if $f(z) = (z - z_1)^2$, condition (15) is not sufficient to prevent the occurrence of errors whatever the value of α_0 .

To ensure that the computed values of α are correct a test must be found such that, if certain conditions are satisfied, the inequality

$$|\alpha| \leq 2\pi - \alpha_0 \quad (17)$$

is true everywhere on the contour. In the following we study one such test.

Consider a circle of radius 1 and assume that the zero (z_0) closest to the contour is located at a distance ρ_0 from the centre of the circle. It is easily seen that the influence of the zero on $|\phi(\theta_l)/\phi(\theta_{l-1})|$ is minimal when the zero is symmetrically

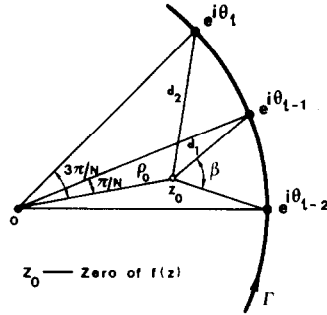


FIG. 1. Geometrical relations for assessing the proximity of a zero relative to Γ .

disposed with respect to two evaluation points on the contour as shown in Fig. 1. From simple geometrical arguments we obtain

$$\left(\frac{d_2}{d_1}\right)^2 = \frac{1 + \rho_0^2 - 2\rho_0 \cos(3\pi/N)}{1 + \rho_0^2 - 2\rho_0 \cos(\pi/N)}, \quad (18)$$

$$\operatorname{tg}\left(\frac{\beta}{2}\right) = \frac{\sin(\pi/N)}{\cos(\pi/N) - \rho_0}. \quad (19)$$

Moreover, for a zero of order k located at the point $z = z_0$ we have

$$\left|\frac{\phi_l}{\phi_{l-1}}\right| \simeq \left(\frac{d_2}{d_1}\right)^k, \quad (20)$$

$$\arg\left[\frac{\phi_{l-1}}{\phi_{l-2}}\right] \simeq k\beta, \quad (21)$$

where $\phi_l = \phi(\theta_l)$.

To begin with we assume $k = 2$. By imposing the condition $2\beta \leq 2\pi - \alpha_0$ we ensure that the error condition (16) is never reached. Since d_2/d_1 is an increasing function of β for a fixed N , a condition on d_2/d_1 equivalent to $2\beta \leq 2\pi - \alpha_0$ is provided by formulas (18) and (19). Thus, through (20), the value of $|\phi_l/\phi_{l-1}|$ gives a test on the violation of condition (17). It can be easily verified that stronger restrictions on $\tilde{\alpha}_l$ lead to weaker restrictions on $|\phi_l/\phi_{l-1}|$ and, conversely, weaker restrictions on $\tilde{\alpha}_l$ lead to stronger restrictions on $|\phi_l/\phi_{l-1}|$. A number of numerical experiments have convinced us that a good choice is $\alpha_0 = 3\pi/4$ which corresponds to $(d_2/d_1)^2 < 6.1$ for any $N \geq 32$. It can be shown that for zeros of order 3 this inequality still ensures that (17) is satisfied for $\alpha_0 = 3\pi/4$. From the foregoing considerations we define the following test on the computed values of ϕ_l/ϕ_{l-1} .

Test. If for all l ϕ_l is such that

$$(i) \quad |\arg(\phi_l/\phi_{l-1})| < 3\pi/4,$$

$$(ii) \quad 1/6.1 < |\phi_l/\phi_{l-1}| < 6.1$$

the computed value of the index is accepted. If any of the conditions (i) or (ii) is not satisfied the index is recalculated with N replaced by $2N$.

It is to be noted that if we considered the zero closest to the contour to lie in the exterior region, i.e., $\rho_0 > 1$ in Fig. 1, we would arrive at approximately the same results although this is a slightly more favourable configuration.

In Tables II and III we illustrate the application of the above test to functions with a single or a double zero close to the integration contour. In these tables the results in italics correspond either to an erroneous value of the index (column IND) or to situations for which at least one of the test conditions is not satisfied (columns α_M and M_f). Examination of Table III may suggest that the above test is unnecessarily restrictive; two remarks are appropriate here: (i) the case $\rho_0/\rho_r = 0.99$ of Table III is very uncommon as it corresponds to a double zero extremely close to contour; (ii) for $N = 256$ the values of α_M and M_f are near the limits of acceptance by the test.

Besides the cases considered in Tables II and III we have checked the test in many different examples without getting errors in the computation of the index.

TABLE II
Simple Zeros: $f(z) = \sin(\pi z - \pi/4)$

ρ_0/ρ_r		0.95		0.99		
N	IND	α_M	M_f	IND	α_M	M_f
16	0	0.86	99.0	0	0.82	400.1
32	8	0.77	10.3	8	0.74	42.5
64	8	0.39	3.4	8	0.48	12.2
128	8	0.21	1.9	8	0.44	5.3

Note. $\rho_0 = |z_0|$, where z_0 is the zero closest to the contour Γ ($\rho_0 = 3.75$). ρ_r = radius of contour Γ (varies with ρ_0/ρ_r); IND = calculated index; $\alpha_M = (1/\pi) \cdot \text{Max}_l |\arg(\phi_l/\phi_{l-1})|$; $M_f = \text{Max}_l |\phi_l/\phi_{l-1}|$, $|\phi_{l-1}/\phi_l|$.

TABLE III
Double Zeros: $f(z) = \sin^2(\pi z - \pi/4)$

ρ_0/ρ_r		0.95		0.99		
N	IND	α_M	M_f	IND	α_M	M_f
16	-1	0.90	251.7	-1	0.85	5,571.5
32	8	0.86	23.9	6	0.96	556.3
64	8	0.69	5.2	8	0.95	106.3
128	8	0.48	2.6	8	0.88	25.5
256	8	0.28	1.6	8	0.76	7.0
512	8	0.15	1.3	8	0.57	2.8

Note. Notation as indicated in Table II ($\rho_0 = 1.75$).

5. CONCLUSIONS

In the foregoing sections we have proposed new formulas for evaluating the integrals involved in the calculation of the coefficients of the polynomial associated with the given analytic function f in some specified region X of the complex plane. These formulas, which do not require the knowledge of $f'(z)$, were shown to be more accurate than those based on a direct evaluation of the integrals of $z^k f'(z)/f(z)$ ($k = 1, 2, \dots, n$) along the boundary of the region.

As is obvious from an examination of the formulas (5) for the coefficients of the associated polynomial, the only critical point in the application of the method of Delves and Lyness is the determination of the number of zeros of f in the region X . The test proposed in the preceding section for accepting or rejecting the number of zeros computed with a certain number of points N has been found to be entirely reliable. The test is not foolproof but the nature of the problem of the evaluation of the index seems to rule out the possibility of existence of such a test.

Finally, as an indication to the user, we would like to point out that the method of Delves and Lyness works equally well for simple and multiple zeros which is a significant advantage over methods based on sequences derived from Newton's method which will fail to converge in the case of multiple zeros (Gardioli's method [1] is an example of this). This is, in fact, a particular case of the more general situation corresponding to the occurrence of saddle points ($f'(z) = 0$) near the path defined by the sequence used, which may result in one or more zeros being missed. But if a very high accuracy is required (say, greater than 10^{-4}) it may be necessary to use locally a scheme (e.g., Muller's method [11]) to refine the values obtained by the present method since the calculation of the integrals with the required accuracy may be too costly. However, for most physical applications, it is enough to compute the integrals with the least number of points for which the calculated value of the index is correct.

APPENDIX: NOMENCLATURE

a_k	coefficients of $P(z)$
α	argument of $f(z)/f(ze^{-i2\pi/N})$
α_0	maximum value accepted for α
C_k	Fourier coefficients
d_1, d_2	distances of the zero z_0 to two consecutive evaluation points
f	analytic function
$\phi(\theta)$	values of f over Γ or Γ_m in polar coordinates
ϕ_l	computed values of ϕ
$g(\theta)$	computed function in (12) and (13)
Γ	boundary of X
Γ_m	boundary of X_m
i	imaginary unit

I_k	integrals related to f for the region X
L_k	integrals related to f for the subregion X_m
N	number of numerical evaluations of f
n	number of zeros of $P(z)$
$P(z)$	associated polynomial
ρ_Γ	radius of the contour Γ
ρ_0	modulus of z_0
ρ_m	radius of the contour Γ_m
S_k	integrals related to $P(z)$
θ	argument of z
θ_l	argument of z at the evaluation points
X	region of the z plane
X_m	subregion of X
Z	complex variable
Z_0	nearest zero to the contour Γ
Z_j	zeros of f

REFERENCES

1. F. E. GARDIOL, *IEEE Trans. Microwave Theory Techn.* **18** (1970), 601–613.
2. L. M. DELVES AND J. N. LYNES, *Math. Comp.* **21** (1967), 543–560.
3. D. H. LEHMER, *J. Assoc. Comput. Math.* **8** (1961), 151–163.
4. W. PFEIFFER, *J. Comput. Phys.* **33** (1979), 397–404.
5. P. LAMPARIELLO AND R. SORRENTINO, *IEEE Trans. Microwave Theory Tech.* **23** (1975), 457–458.
6. A. BARBOSA, A. F. DOS SANTOS, AND J. FIGANIER, *Proc. IEE* **128**, Pt.H (1981), 243–246.
7. J. D. CALLEN “Absolute and Convective Instabilities of a Magnetized Plasma,” Ph. D. Thesis, MIT Dept. of Nuclear Engineering, 1968.
8. H. G. GARNIR AND J. GOBERT, “Fonctions d’une variable complexe,” p. 83, Dunod, Paris, 1965.
9. P. HENRICI, “Applied and Computational Complex Analysis,” Vol. 1, p. 239, Wiley–Interscience, New York, 1974.
10. G. CAIN, JR., *Comm. ACM.* **9** (1966), 305–306.
11. P. HENRICI, “Elements of Numerical Analysis,” p. 198, Wiley, New York, 1964.