Numerical Differentiation of Analytic Functions
Author(s): J. N. Lyness and C. B. Moler
Source: *SIAM Journal on Numerical Analysis*, Vol. 4, No. 2 (Jun., 1967), pp. 202-210
Published by: Society for Industrial and Applied Mathematics
Stable URL: http://www.jstor.org/stable/2949389
Accessed: 14/12/2009 10:08

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/page/info/about/policies/terms.jsp. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/action/showPublisher?publisherCode=siam.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Society for Industrial and Applied Mathematics is collaborating with JSTOR to digitize, preserve and extend access to *SIAM Journal on Numerical Analysis.*

http://www.jstor.org

SIAM J. NUMER. ANAL.
Vol. 4, No. 2, 1967
Printed in U.S.A.

NUMERICAL DIFFERENTIATION OF ANALYTIC FUNCTIONS*

J. N. LYNESS† AND C. B. MOLER‡

1. Introduction. Many algebraic computer languages now include facilities for the evaluation of functions of a complex variable. Such facilities can be effectively used for numerical differentiation. The method we describe is appropriate for computing the derivatives $f^{(n)}(x)$ of any analytic function which can be evaluated at points in the complex plane near x. Such a function might be a complicated rational combination of elementary functions of x which is difficult to differentiate analytically.

In this paper we derive several formulas, any of which is suitable for evaluating the nth derivative of a complex analytic function at a point in terms of function evaluations at neighboring points. For simplicity we assume throughout this paper that the derivatives are to be evaluated at the origin. In this Introduction we state one formula; in succeeding sections we prove this and related formulas, discuss their degree and the error in their application, and give an example.

A trapezoidal rule sum operator $R^{[m,1]}$ may be defined as follows:

$$(1.1)\quad R^{[m,1]}g(t) = \frac{1}{m}\left[\frac{1}{2}g(0) + g\left(\frac{1}{m}\right) + \cdots + g\left(\frac{m-1}{m}\right) + \frac{1}{2}g(1)\right].$$

In (2.12) below we define the Möbius numbers μ_n, $n = 1, 2, \cdots$, which have the values $+1$, 0, or -1. In terms of

$$(1.2)\qquad g(r;t) = g(t) = \operatorname{Re} f(re^{2\pi it}),$$

one of our results may be stated as follows.

THEOREM 1. *If $f(x)$ is a real function of x and $f(z)$ is analytic in a neighborhood that contains the circle $|z| = r$, then*

$$(1.3)\qquad a_n = \frac{f^{(n)}(0)}{n!} = \frac{1}{r^n}\sum_{m=1}^{\infty}\mu_m[R^{[mn,1]}g(t) - f(0)].$$

In other words, the nth derivative of f can be computed from the trapezoidal rule operating on the real part of f. The only complex arithmetic required is in the computation of this real part.

This result is based on Cauchy's theorem, which expresses the nth deriva-

* Received by the editors August 15, 1966, and in revised form December 20, 1966. This research was sponsored in part by the United States Atomic Energy Commission under contract with the Union Carbide Corporation.

† Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37830. On leave from the University of New South Wales, Sidney, Australia.

‡ Department of Mathematics, University of Michigan, Ann Arbor, Michigan.

tive of an analytic function in terms of a contour integral. This integral contains an oscillating component and may be evaluated using the Poisson summation formula and the Möbius inversion of series.

Since this method is based on numerical quadrature, it does not show the sensitivity to roundoff error in the function evaluations that is characteristic of finite difference methods.

2. Proof of Theorem 1. Cauchy's theorem may be stated as

$$a_n \equiv \frac{f^{(n)}(0)}{n!} = \frac{1}{2\pi i}\int_c \frac{f(z)}{z^{n+1}}\,dz, \qquad n = 0, 1, 2, \cdots, \tag{2.1}$$

where $f(z)$ is analytic in a neighborhood which includes a closed contour C which surrounds the origin.

If we choose C to be the circle $|z| = r$ and use variable t defined by

$$z = re^{2\pi it}, \tag{2.2}$$

we find by direct substitution

$$a_0 = f(0) = \int_0^1 f(re^{2\pi it})\,dt, \tag{2.3}$$

$$a_n = \frac{1}{r^n}\int_0^1 f(re^{2\pi it})e^{-2\pi int}\,dt, \qquad n = 1, 2, 3, \cdots. \tag{2.4}$$

Using (2.3) with $f(z)$ replaced by $z^n f(z)$, we find

$$0 = \frac{1}{r^n}\int_0^1 f(re^{2\pi it})e^{2\pi int}\,dt, \qquad n = 1, 2, 3, \cdots. \tag{2.5}$$

Equations (2.4) and (2.5) may be combined to give either of the two forms:

$$a_n = \frac{2}{r^n}\int_0^1 f(re^{2\pi it})\cos 2\pi nt\,dt, \tag{2.6}$$

$$a_n = \frac{-2i}{r^n}\int_0^1 f(re^{2\pi it})\sin 2\pi nt\,dt. \tag{2.7}$$

The Poisson summation formula in its finite form [5] connects the trapezoidal rule sum operator (1.1) with the Fourier coefficients (2.6). It states

$$\begin{aligned} R^{[n,1]}f(re^{2\pi it}) &= \sum_{k=-\infty}^{\infty}\int_0^1 f(re^{2\pi it})e^{2\pi iknt}\,dt \\ &= \int_0^1 f(re^{2\pi it})\,dt + 2\sum_{k=1}^{\infty}\int_0^1 f(re^{2\pi it})\cos 2\pi knt\,dt. \end{aligned} \tag{2.8}$$

If we denote by b_n the error in the n-point trapezoidal rule:

$$b_n = \left[R^{[n,1]}f(re^{2\pi it}) - \int_0^1 f(re^{2\pi it})\,dt\right] \tag{2.9}$$
$$= R^{[n,1]}f(re^{2\pi it}) - f(0),$$

then (2.8) may be written in the form

$$b_n = r^n a_n + r^{2n}a_{2n} + r^{3n}a_{3n} + \cdots, \quad n = 1, 2, 3, \cdots. \tag{2.10}$$

The inversion of a set of equations of this type is a familiar exercise in elementary number theory. The result is

$$r^n a_n = \mu_1 b_n + \mu_2 b_{2n} + \mu_3 b_{3n} + \cdots, \quad n = 1, 2, 3, \cdots, \tag{2.11}$$

the μ_n being the nth Möbius number defined by (see, for example, [4])

$$\begin{aligned} \mu_1 &= 1, \\ \mu_i &= (-1)^r \quad \text{if } i \text{ is a product of } r \text{ distinct prime numbers,} \\ \mu_i &= 0 \quad \text{otherwise.} \end{aligned} \tag{2.12}$$

The values of the first fifteen Möbius numbers are given here together with the values λ_k needed later on.

$$\begin{aligned} k &= 1, \ 2, \ 3, 4, \ 5, \ 6, \ 7, 8, 9, \ 10, \ 11, \ 12, \ 13, \ 14, \ 15, \\ \mu_k &= +1, -1, -1, 0, -1, +1, -1, 0, 0, +1, -1, \ 0, -1, +1, +1, \\ \lambda_k &= +1, +1, -1, 2, -1, -1, -1, 4, 0, -1, -1, -2, -1, -1, +1. \end{aligned}$$

If $f(x)$ is a real function of x, it follows that a_n is real. If we substitute into the real part of (2.11) the expression for b_n given by (2.9) and a_n by (2.1), we find (1.3), thus establishing Theorem 1.

3. Other theorems of a similar nature. There are several generalizations and modifications of the above derivation which lead to formulas of a nature similar to that of Theorem 1. In this section we outline these derivations and state corresponding results.

We introduce a set of rule sum operators in analogy to (1.1) above. These are "off set" trapezoidal rules,

$$R^{[n,\alpha]}g(t) = \frac{1}{n}\sum_{j=1}^{n} g\left(\frac{2j + \alpha - 1}{2n}\right), \qquad |\alpha| < 1. \tag{3.1}$$

The Fourier series corresponding to one of these forms a generalization of the Poisson summation formula (see [2]), namely,

$$R^{[n,\alpha]}g(t) = \int_0^1 g(t)\,dt + \sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} e^{-\pi i(\alpha-1)k}\int_0^1 g(t)e^{2\pi iknt}\,dt. \tag{3.2}$$

We are interested here in cases in which $\alpha = 0, \pm\frac{1}{2}, 1$. We define in analogy to (2.9) above,

$$
\begin{aligned}
b_n^{[0]} &= R^{[n,0]}f(re^{2\pi it}) - \int_0^1 f(re^{2\pi it})\,dt, \\
b_n^{[1]} &= b_n = R^{[n,1]}f(re^{2\pi it}) - \int_0^1 f(re^{2\pi it})\,dt, \\
b_n^{[2]} &= \tfrac{1}{2}(b_n^{[1]} - b_n^{[0]}), \\
ib_n^{[3]} &= \tfrac{1}{2}(R^{[n,-1/2]}f(re^{2\pi it}) - R^{[n,1/2]}f(re^{2\pi it})).
\end{aligned} \tag{3.3}
$$

Substituting for a_n given by (2.3) or (2.4) into (3.3), we find respectively

$$
\begin{aligned}
b_n^{[0]} &= -r^n a_n + r^{2n}a_{2n} - r^{3n}a_{3n} + \cdots, \\
b_n^{[1]} &= r^n a_n + r^{2n}a_{2n} + r^{3n}a_{3n} + \cdots, \\
b_n^{[2]} &= r^n a_n + r^{3n}a_{3n} + r^{5n}a_{5n} + \cdots, \\
b_n^{[3]} &= r^n a_n - r^{3n}a_{3n} + r^{5n}a_{5n} - \cdots, \qquad n = 1, 2, 3, \cdots.
\end{aligned} \tag{3.4}
$$

All these equations may be inverted in a similar way. The results are given in the following theorem.

THEOREM 2. *Under the hypotheses of Theorem* 1,

$$
\begin{aligned}
r^n a_n &= -\sum_{m=1}^{\infty} \lambda_m b_{mn}^{[0]}, \\
r^n a_n &= \sum_{m=1}^{\infty} \mu_m b_{mn}^{[1]}, \\
r^n a_n &= \sum_{m=1}^{\infty} \mu_{2m-1} b_{(2m-1)n}^{[2]}, \\
r^n a_n &= i\sum_{m=1}^{\infty} (-1)^m \mu_{2m-1} b_{(2m-1)n}^{[3]},
\end{aligned} \tag{3.5}
$$

where the set of numbers λ_m *are defined by*

$$
\begin{aligned}
\lambda_m &= \mu_m, \qquad m \text{ odd}, \\
\lambda_m &= 2^{\alpha-1}\mu_{2n+1}, \qquad m = 2^\alpha(2n+1), \qquad \alpha \geqq 1.
\end{aligned} \tag{3.6}
$$

Any of the set (3.5) may be used to compute a_n. If $f(x)$ is a real function of x, the first three involve only the real part of $f(z)$ and the fourth only the imaginary part of $f(z)$. Whether or not $f(x)$ is a real function of x, the first two involve the function evaluation $f(0)$ while the second two do not require this function evaluation.

4. Convergence and degree. The rate of convergence of the series in Theorem 2 is indicated by the following theorem.

THEOREM 3. *For all p,*

$$\lim_{m\to\infty} b_m^{[k]}\, m^p = 0, \qquad k = 0, 1, 2, 3.$$

In other words the sequence $b_m^{[k]}$ approaches zero faster than any power of m. We present a proof for $k = 1$. The proof for the other values of k is similar.

Proof. The Euler Maclaurin summation formula states:

$$R^{[m,1]}h(t) - \int_0^1 h(t)\,dt = 2\sum_{s=1}^{p-1} (-1)^{s+1}\zeta(2s)\,\frac{h^{(2s-1)}(1) - h^{(2s-1)}(0)}{(2\pi m)^{2s}} + \frac{2(-1)^p}{(2\pi m)^{2p}} \int_0^1 h^{(2p)}(t) \sum_{r=1}^{\infty} \frac{1 - \cos 2\pi rmt}{r^{2p}}\,dt, \tag{4.1}$$

where $h(t)$ and all its derivatives are continuous in the interval $0 \leqq t \leqq 1$. Since $f(z)$ is analytic in a domain which includes the circle $|z| = r$, it follows that

$$h(t) = f(re^{2\pi it})$$

satisfies the above condition. In addition $h(t)$ is periodic and analytic. Thus,

$$h^{(2s-1)}(1) - h^{(2s-1)}(0) = 0 \tag{4.2}$$

and the $(2p)$th derivative of $h(t)$ is bounded,

$$|h^{(2p)}(t)| \leqq M_{2p}. \tag{4.3}$$

If we use the inequality

$$\left| \sum_{r=1}^{\infty} \frac{1 - \cos 2\pi rmt}{r^{2p}} \right| \leqq 2\zeta(2p) \tag{4.4}$$

and the definition (3.3), we find without difficulty that

$$|b_m^{[1]}| \leqq \frac{4M_{2p}\,\zeta(2p)}{(2\pi m)^{2p}}. \tag{4.5}$$

The theorem follows immediately from this.

The formulas of Theorem 2 express the nth derivative as an infinite sum of terms, each of which is a sum of function evaluations. In practice this infinite sum is approximated by the sum of the first M terms, and M is determined by the size of the terms and the apparent convergence rate of the series. The connection between M and the degree of the formula is indicated by the following theorem.

THEOREM 4. *The formula*

$$a_n = \frac{1}{r^n} \sum_{m=1}^{M} \mu_m b_{mn}^{[1]} \tag{4.6}$$

and the corresponding finite sums for $k = 0, 2, 3$ *are exact if* $f(z)$ *is a polynomial of degree* d, *where*

$$\begin{aligned} d &< (M+1)n, \qquad k = 0, 1, \\ d &< (2M+1)n, \qquad k = 2, 3. \end{aligned} \tag{4.7}$$

Proof. If $f(x)$ is a polynomial of degree d, then

$$a_n = 0, \qquad n > d, \tag{4.8}$$

and using (3.4),

$$b_n^{[k]} = 0, \qquad n > d. \tag{4.9}$$

It follows that if d satisfies (4.7) the sum given in (4.6) is identical with the sum in (3.5) which, by Theorem 2, is exact.

5. A roundoff error estimate. From a practical standpoint, one of the most important features of these methods is that a simple estimate of the roundoff error is available at an early stage in the calculation. In this section we derive this estimate and indicate how it may be used to choose a value for the parameter r suitable to the required accuracy of the particular calculation.

We assume that the calculation is carried out using floating point arithmetic. We denote by ϵ the maximum relative roundoff error in the arithmetic operations and in the function evaluations. In the numerical examples below, the computer used 36-bit floating point significant arithmetic, and

$$\epsilon = 0.5 \times 10^{-10}. \tag{5.1}$$

As an illustration of these methods we consider the calculation of the fifth derivative $f^{(5)}(0)$ of the function

$$f(x) = \frac{e^x}{\sin^3 x + \cos^3 x}. \tag{5.2}$$

(See Table 1.) This example satisfies our requirements of being tedious to differentiate analytically, but easy to compute for complex arguments. It is clear that the value of the required derivative is integer. Using a symbolic algebraic processor [1]—certainly the most convenient way to solve this particular problem—we find

$$f^{(5)}(0) = -164. \tag{5.3}$$

Conventional methods of numerical differentiation using divided differences or methods using the Neville algorithm [3] are based implicitly on formulas of the type

$$f^{(n)}(0) = \sum a_i f(x_i), \tag{5.4}$$

TABLE 1

Calculation of $f^{(5)}(0)$ where $f(x)$ is given by (5.2)

(a) *The value of ϵ is given by* (5.1).

(b) *In each case, the value of δ_1 is available after the first three function evaluations.*

(c) *The series is terminated after the Mth term, M being the first integer for which $|b_{Mn}| < \epsilon G_M$ and $\mu_M \neq 0$. In each case the sequence $|b_{Mn}|$ is strongly monotonic decreasing at this stage.*

(d) *The quantity ν is the number of times the subroutine for evaluating $f(z)$ was entered in our particular code. This is not the minimum possible, since we found it convenient to evaluate the function at a point each time this quantity was required. However, we did make use of the fact that $g(t)$ is symmetric about $t = \frac{1}{2}$ which approximately halves the number of function evaluations required.*

r	M	S_M	δ_1	δ_M	$f^{(5)}(0)$	$\delta_M \lvert f^{(5)}(0) \rvert$	ν
0.1	3	$-1{\cdot}3666605809 \times 10^{-5}$	$4{\cdot}10 \times 10^{-6}$	$4{\cdot}10 \times 10^{-6}$	$-163{\cdot}99926{,}973$	$0{\cdot}00067{,}246$	16
0.4	7	$-1{\cdot}3994666602 \times 10^{-2}$	$6{\cdot}52 \times 10^{-9}$	$6{\cdot}34 \times 10^{-9}$	$-163{\cdot}99999{,}925$	$0{\cdot}00000{,}104$	59
0.7	37	$-2{\cdot}2969566684 \times 10^{-1}$	$8{\cdot}99 \times 10^{-10}$	$6{\cdot}13 \times 10^{-10}$	$-164{\cdot}00000{,}013$	$0{\cdot}00000{,}010$	1097

in which the constants a_i vary considerably in magnitude and are of different signs. This causes small errors in the function values $f(x_i)$ to be amplified in a manner which is difficult to predict quantitatively. In the above example, a fifth degree polynomial through computed function values at six points evenly spaced in the interval $[-0.1, 0.1]$ gives

$$f^{(5)}(0) = -168.5 \cdots . \tag{5.5}$$

An exhaustive experiment using several dozen different point distributions and different intervals produced results both more and less accurate than this. The closest result was correct to five significant figures but some results were even of the wrong sign.

The methods described in this paper involve formulas of type (5.4), all the a_i having the same magnitude. Specifically the steps involved in using (1.3) are

$$R^{[mn,1]}g(t) = \frac{1}{mn}\sum_{j=1}^{mn} g(t_j), \qquad t_j = \frac{j}{mn}, \tag{5.6}$$

$$b_{mn} = R^{[mn,1]}g(t) - f(0), \tag{5.7}$$

$$S_M = \sum_{m=1}^{M} \mu_m b_{mn}, \qquad |\mu_m| = 1 \text{ or } 0, \tag{5.8}$$

$$\frac{f^{(n)}(0)}{n!} = a_n \doteq \frac{S_M}{r^n}. \tag{5.9}$$

While each of these steps involves some roundoff error, it is clear that in (5.6), (5.7) and (5.8) there is no undue amplification of the *absolute*

value of the errors in $g(t_j)$. In fact, the averaging process in (5.6) may damp these errors down. Consequently, the computed S_M may be expected to have an absolute error of the order of ϵG_M, where $G_M = G_M(r)$ is the largest function value occurring in the calculation of S_M, that is,

$$G_M = \max \{|f(0)|, |g(t_j)|\}, \qquad t_j = j/mn, \ 0 \leqq m \leqq M. \tag{5.10}$$

The final step (5.9) retains the *relative* accuracy. Thus a guide to the relative accuracy in the final result is given by δ_M, where

$$\frac{\Delta f^{(n)}(0)}{f^{(n)}(0)} \doteq \delta_M = \frac{\epsilon G_M}{S_M} \doteq \frac{\epsilon G(r)}{r^n a_n}, \tag{5.11}$$

where

$$G(r) = \max |g(t)| = \lim_{M \to \infty} G_M(r). \tag{5.12}$$

The possibly large size of this relative error is a consequence of (5.6) and (5.7). If $|f(0)|$ is small or zero, then (5.6) involves numbers $g(t_j)$ of different sign. If $|f(0)|$ is large, (5.7) may involve taking the difference of two nearly equal numbers. In either case, relative accuracy is lost, although absolute accuracy is retained.

The value δ_M is in no rigorous sense a bound. The actual accuracy may be less than δ_M if roundoff errors unexpectedly cancel. Or one may have a slightly larger error, particularly if M is large. For example, if $f(z)$ involves a large nearly constant component, then the accumulation of the sum in (5.6) may be an additional significant source of error. This could be avoided at little expense by using double precision accumulations.

The estimate δ_M is an a posteriori estimate. Remembering that δ_M is not required at all accurately, but that only its rough magnitude is needed, the term δ_1 is usually a sufficiently accurate approximation to δ_M and is therefore an estimate of the roundoff error in the final result.

The advantage in using δ_1 is that it provides an *early* indication of whether the chosen value of the parameter r is suitable in view of the required accuracy. If δ_1 indicates an unacceptably large error, the calculation may be recommenced at once using a larger value of r. Since the series converges more slowly if r is larger, the coding of the problem may be arranged to use ultimately as small a value of r as is convenient and also consistent with the acceptable roundoff error.

6. Conclusion. It seems inappropriate to compare in any detail the methods given here with standard methods requiring function evaluations with real argument only. From the point of view of the numerical analyst these methods show that, once complex arguments are allowed, the principal difficulties encountered in numerical differentiation simply disappear. Theo-

rem 4 indicates that it is trivial to invent methods of any degree and the discussion of §5 shows that the roundoff error may be controlled in a straightforward manner. The methods given here are not minimal in the sense relating to the degree of the polynomial approximation and for this reason may well be unacceptable to the more conventional numerical analyst.

For the practical computer user with a function to differentiate, these methods are quite attractive. They are based on simple formulas, are easy to code, converge rapidly and, perhaps most important, include a useful error estimate.

7. Acknowledgment. Both of the authors are happy to acknowledge help and support received during independent visits to the Institut für Angewandte Mathematik, Eidgenössische Technische Hochschule, Zürich, Switzerland, where most of this work was carried out.

REFERENCES

[1] M. ENGELI, *Design and implementation of an algebraic processor*, Privately circulated report, Eidgenössische Technische Hochschule, Zürich, 1966.

[2] J. N. LYNESS, *The calculation of Fourier coefficients*, this Journal, 4 (1967), pp. 301–315.

[3] J. N. LYNESS AND C. B. MOLER, *Van der Monde systems and numerical differentiation*, Numer. Math., 8 (1966), pp. 458–464.

[4] G. PÓLYA AND G. SZEGÖ, *Aufgaben und Lehrsätze aus der Analysis*, vol. 2, Springer-Verlag, Berlin, 1954.

[5] E. C. TITCHMARSH, *Introduction to the Theory of Fourier Integrals*, Oxford University Press, Oxford, 1948.