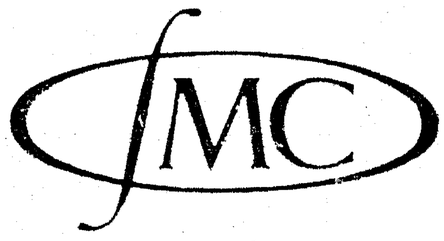
STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

MR 36

Numerical Efficiency Profile Functions

P. Wynn



1962

KONINKL. NEDERL. AKADEMIE VAN WETENSCHAPPEN – AMSTERDAM
Reprinted from Proceedings, Series A, 65, No. 1 and Indag. Math., 24, No. 1, 1962

MATHEMATICS

NUMERICAL EFFICIENCY PROFILE FUNCTIONS [1])

BY

P. WYNN

(Communicated by Prof. A. VAN WIJNGAARDEN at the meeting of September 30, 1961)

The numerical computation of a mathematical function by use of a recursive process may only be carried out if the number of times which the process must be repeated in order to achieve a prerequired accuracy, for given values of the arguments, is known. A useful device for the investigation of the numerical efficiency of a recursive process for computing a function of one argument, is a two argument table of the following form:

TABLE I

	s_1	s_2	.	.	.	s_{ha}
z_1	$n_{1,1}$	$n_{1,2}$	.	.	.	$n_{1,ha}$
z_2	$n_{2,1}$	$n_{2,2}$	.	.	.	$n_{2,ha}$
.	.	.				.
.	.	.				.
z_{ka}	$n_{ka,1}$	$n_{ka,2}$	.	.	.	$n_{ka,ha}$

This table indicates the necessary number $n_{r,r'}$ of repetitions of the process for which a relative error less than or equal to $(\frac{1}{2})p^{-s}$ (where p is a given radix) when the argument has the value z, may be attained.

The two arguments are derived from the sequence of s values $s_1, s_2, \ldots, s_{ha}$ and the sequence of z values $z_1, z_2, \ldots, z_{ka}$. It is usually true that the arguments may be chosen such that z_r, $r = 1, \ldots, ka$ and $s_{r'}$, $r' = 1, \ldots, ha$ are monotonic sequences, and $n_{r,r'}$ is a monotonic function which increases with both r and r'. For the consideration of processes for computing functions of more than one argument, a set of such tables may be given.

For the construction of a subroutine for computing a function by use of a power series or continued fraction expansion, there must be provided a) an auxiliary subroutine which computes the coefficients and b) a table having the form of Table I, or a set of such tables.

If the relative error in the computation of the function value must not exceed $\frac{1}{2}p^{-s}$, the argument value is z', $z_r < z' \leqslant z_{r+1}$ and $s_{r'} < s' \leqslant s_{r'+1}$ then the index of the required partial sum or convergent is $n_{r+1,r'+1}$.

The requirement a), however exacting, is unavoidable. The requirement b) may be dispensed with if, in the case of the computation of a function

[1]) Communication MR 36 of the Computation Department of the Mathematical Centre at Amsterdam.

of one argument, a numerical efficiency profile function $n(z, s)$ of continuous variables s and z may be given, for which

$$n(z_r, s_{r'}) \geqslant n_{r,r'} \qquad r = 1, 2, \ldots, ka; \qquad r' = 1, 2, \ldots, ha.$$

The use of this function may result in the computation of a partial sum or convergent of higher order than is necessary, but this, if the profile function itself is easily computed, may easily be outweighed by the advantage of being able to dispense with the input of sets of tables of numerical data. Of course the profile function should be so chosen that the quantities $n(z_r, s_{r'}) - n_{r,r'}$ are as small as possible. The general statement of the problem of constructing profile functions which occur in the computation of functions of more than one argument, is obvious by extension of the preceding considerations.

It is possible to verse the problem as a linear programming problem. It is assumed that

$$n(z, s) = \sum_{u=0}^{m} b_{u'} \phi_u(z, s). \tag{1}$$

There then result the ha $\times$ ka linear inequalities

$$n_{r,r'} \leqslant \sum_{u=0}^{m} b_{u'}\, \phi_u(z_r, s_{r'}) \qquad r = 1, 2, \ldots, ka;\ r' = 1, \ldots, ha \tag{2}$$

and the linear function to be maximised is provided by the condition that the double integral

$$\sum_{u=0}^{m} b_{u'} \cdot \int_{s_1}^{s_{ha}} \int_{z_1}^{z_{ka}} \phi_u(z, s)\, dzds \tag{3}$$

is to be a minimum. (A cosmetic generalisation is possible at this point. It is formally possible to introduce a weight function $f(z, s)$ inside the integral sign in (3). This may be of use, for example, if it is known that the subroutine is to be used far more over certain ranges of z and s than for others; $f(z, s)$ would then be an approximation to a frequency function.)

If, however, the function $n(z, s)$ is taken to be

$$n(z, s) = \sum_{h=0}^{h=nd} \sum_{u=0}^{h} b_{h,u} z^{h-u} s^u \tag{4}$$

with $nd = 3$, $ha = 11$, $ka = 6$ (referring to Tables given in [1]), and the determination of the coefficients in (4) is embarked upon as a straightforward linear programming problem (including the introduction of artificial variables), it may easily be shown that repeated operation upon an array of some 73 $(72 + 20 + 72 + 1) = 12{,}045$ quantities is required, and that the construction of profile functions for functions of more than one argument require operation upon an astronomical number of quantities. Nevertheless certain economies can be effected, as will become apparent by considering the following exposition. (It is assumed that the reader

is fully conversant with the linear programming problem and the simplex method, as described inter alia in [2] and [3]).

i) The numerical efficiency profile function is taken to be

$$P - \sum_{h=0}^{h=nd} \sum_{u=0}^{h} b_{h,u}\, z^{h-u} s^{u} \tag{5}$$

where $P > \max n_{r,r'}$. By means of this artifice the introduction of artificial variables as a basic feasible solution is dispensed with.

ii) Since it is not a priori known if the coefficients in the expression (5) are positive or negative, (5) implies the existence of $(nd+1)(nd+2) = m$ unknown positive quantities $b_{h,u\ \mathrm{pos}}$, $b_{h,u\ \mathrm{neg}}$ where

$$b_{h,u} = b_{h,u\ \mathrm{pos}} - b_{h,u\ \mathrm{neg}}.$$

iii) There are $N = ha \times ka$ inequalities of the form

$$P - \sum_{h=0}^{h=nd} \sum_{u=0}^{h} b_{h,u}\, z_r{}^{h-u} s_{r'}{}^{u} \geqslant n_{r,r'} \qquad r = 1, 2, \ldots, ka;\ r' = 1, 2, \ldots, ha.$$

Thus the variables occurring in the linear programming problem are the set $x_u\ u = 1, \ldots, m$ where

$$x_1 = -b_{0,0\ \mathrm{pos}},\ x_2 = P + b_{0,0\ \mathrm{neg}},\ \ldots,\ x_m = b_{nd,nd\ \mathrm{neg}}$$

and the residual variables $x_{m+s}\ s = 1, \ldots, N$ given by

$$P - \sum_{h=0}^{h=nd} \sum_{n=0}^{h} b_{h,u}\, z_r{}^{h-u} s_{r'}{}^{u} + x_{m+(r'-1)\,ka+r} \qquad \begin{matrix} r = 1, \ldots, ka \\ r' = 1, \ldots, ha \end{matrix}. \tag{6}$$

iv) The linear function to be maximised is $\sum_{i=n}^{n} c_i x_i$

$$c_1 = \frac{(z_{ka} - z_1)\,(s_{ha} - s_1)}{1.1} \qquad c_2 = -c_1$$

$$c_3 = \frac{(z_{ka}^2 - z_1{}^2)\,(s_{ha} - s_1)}{2.1} \qquad c_4 = -c_3$$

$$\ldots \qquad \ldots$$

$$c_{m-1} = \frac{(z_{ka} - z_1)\,(s_{ha}^{nd+1} - s_1{}^{nd+1})}{1.\ (nd+1)} \qquad c_m = -c_{m-1}.$$

The coefficients $c_{m+u}\ u = 1, \ldots, N$ corresponding to the residual variables x_{m+u}, are zero.

v) The simplex method requires repeated operation upon quantities in an $N \times (m+N)$ array $a_{i,j}$. Initially the first m columns of this array are as follows

$$
\begin{array}{cccccccccccc}
1 & -1 & z_1 & -z_1 & s_1 & -s_1 & \cdot\ \cdot\ \cdot & z_1 s_1^{nd-1} & -z_1 s_1^{nd-1} & s_1^{nd} & -s_1^{nd} \\
1 & -1 & z_2 & -z_2 & s_1 & -s_1 & \cdot\ \cdot\ \cdot & z_2 s_1^{nd-1} & -z_2 s_1^{nd-1} & s_1^{nd} & -s_1^{nd} \\
1 & -1 & z_3 & -z_3 & s_1 & -s_1 & \cdot\ \cdot\ \cdot & z_3 s_1^{nd-1} & -z_3 s_1^{nd-1} & s_1^{nd} & -s_1^{nd} \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot\ \cdot\ \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & & \cdot & \cdot & \cdot & \cdot \\
1 & -1 & z_{ka} & -z_{ka} & s_1 & -s_1 & \cdot\ \cdot\ \cdot & z_{ka} s_1^{nd-1} & -z_{ka} s_1^{nd-1} & s_1^{nd} & -s_1^{nd} \\
1 & -1 & z_1 & -z_1 & s_2 & -s_2 & \cdot\ \cdot\ \cdot & z_1 s_2^{nd-1} & -z_1 s_2^{nd-1} & s_2^{nd} & -s_2^{nd} \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & & \cdot & \cdot & \cdot & \cdot \\
1 & -1 & z_{ka} & -z_{ka} & s_{ha} & -s_{ha} & \cdot\ \cdot\ \cdot & z_{ka} s_{ha}^{nd-1} & -z_{ka} s_{ha}^{nd-1} & s_{ha}^{nd} & -s_{ha}^{nd}
\end{array}
$$

(These may easily be built up by forming sets of $2(h+1)$ columns $h=2.\ 3, \ldots, nd$ by partitioned multiplication.)

The further N rows of the array (corresponding to the variables x_{m+u} $u=1, \ldots, N$ which at the start are non-zero) possess but one non-zero element.

vi) The basic feasible solution is taken to be

$$x_u = 0 \qquad u = 1, 2, \ldots, m$$

$$x_{m+u} = P - n_{r,r'} \qquad u = (r'-1)ka + r \qquad r = 1, \ldots, ka;\ r' = 1, \ldots, ha$$

vii) As the solution proceeds N of the variables x_u will be non-zero; their values are appended as a column to the a-array, their indices are appended as a further column.

viii) The simplex method proceeds by the recursive exchange of variables between the zero set and the non-zero set. The index of the variable to be discarded from the zero set is determined by selecting the most negative of a set of m quantities (each of which corresponds to one of the zero variables). These quantities are appended as a further row to the a-array (their initial values are $-c_u$ $u=1, 2, \ldots, m$). The indices of the zero variables are appended as a further row to the a-array.

ix) Principal interest attaches to the numerical values of the variables x_u $u=1, 2, \ldots, m$. In this instance it is convenient to adjoin a further row, the u^{th} member ($u=1, 2, \ldots, m$) of which indicates the row number of the extended a-array which contains the value of the variable x_u, should this be non-zero. In this way a tag is kept upon the variables x_u, $u=1, 2, \ldots, m$ during the course of the computation.

The extended array just described will be referred to as the f-array. The computation now runs as follows:

Prepare

Compute

					Comment
$f_{i,j}$	$i = 1(1)N,$	$j =$	$1(1)m$		v)
$f_{N+1,j}$	$= -c_j$	$j =$	$1(1)m$		viii)
$f_{i,m+1}$	$= P - n_{r,r'}$	$i =$	$1(1)N$		(vi) + vii)
$f_{i,m+2}$	$= m + i$	$i =$	$1(1)N$		vii)
$f_{N+2,j}$	$= j$	$j =$	$1(1)m$		vi) + viii)
$f_{N+3,j}$	$= 0$	$j =$	$1(1)m$		ix)

Simplex Algorithm

$$e = 0;\ v = 0;$$

Determine e from

$$f_{N+1,e} = \max \text{ neg } f_{N+1,j} \quad j = 1(1)m$$

e zero? — yes → Output

|

no

↓

Determine v from

$$\left(\frac{f_{v,m+1}}{f_{v,e}}\right) = \min \text{ pos} \left(\frac{f_{i,m+1}}{f_{i,e}}\right) \qquad i = 1(1)N$$

v zero? — yes → Process chronically unstable.

| Repeat using greater precision.

no

↓

new $f_{v,j} = \left(\frac{\text{old } f_{v,j}}{f_{v,e}}\right)$ $\quad j = 1(1)m+1,\ j \neq e;$ new $f_{v,e} = 1/(\text{old } f_{v,e});$

new $f_{i,j} = (\text{old } f_{i,j}) - f_{v,j} \times f_{i,e}$ $\quad i = 1(1)N+1,\ i \neq v;\ j = 1(1)m+1,\ j \neq e;$

new $f_{i,e} = -(\text{old } f_{i,e}) \times f_{v,e}$ $\quad i = 1(1)N+1,\ i \neq v;$

if $\text{ind} = f_{N+2,e} \leqslant m$ then $f_{N+3,e} = v$ ix)

if $\text{ind}' = f_{v,m+2} \leqslant m$ then $f_{N+3,\,\text{ind}'} = 0$ ix)

Interchange $f_{N+2,e}$ and $f_{v,m+2}$ vii) + viii)

Repeat Simplex Algorithm

Output

pos $j = f_{N+3,2j-1}$, neg $j = f_{N+3,2j}$

Print $b_j = f_{\text{pos } j,m+1} - f_{\text{neg } j,m+1}$ $\quad j = 1(1)m/2$ ii)

The calculations may be finally checked (and indeed at any intermediate stage) by printing out the values of the variables x_u $u = 1(1)m+N$ and evaluating equations (5) and (6).

By carrying out the computations in the manner described above the array referred to at the beginning of section has been reduced to one of $(20+2)\,(72+3) = 1650$ quantities.

If the profile function is taken to be a plane (i.e. $nd=1$ in equation (4) and the given values of $n_{r,r'}$ lie on a wholly convex surface, an upper limit to the number of steps in the simplex algorithm may be given. At any stage of the process (other than the first two) the values of the residual variables at three neighbouring points in the $z-s$ plane are zero. The computation proceeds by removing one of the points from the group of three, and choosing another which is neighbour to the remaining two. Indeed the progress of the computation may in this case be regarded as a slow waltz by a three-legged Thing on the $z-s$ plane. The most prolonged calculation possible occurs when the waltz takes place from one corner of the plane to the opposite corner, that is, taking $N+m$ steps. (Note: the values of $n_{r,r'}$ displayed in Table II (to be given) do not, as is revealed by inspection of Table III, lie on a surface which is wholly convex in the sense described).

The progress of the waltz may, if desired, be observed by printing out and inspecting the quantities $f_{N+2,j}$ $j=1(1)m$ at each stage of the Simplex Algorithm.

It does not appear that any such upper limit to the extent of the computation may be given when the profile function is a surface of higher degree than one, and indeed in this case the progress of the computation with regard to the positioning of the successive zero residuals recalls one of the wilder moments in the Patagonian Rhumba.

If the greatest value of $n_{r,r'}$ occurs at the point $z=z_{ka}$, $s=s_{ha}$ then it is quite clear that the step which reduces the value of the integral (3) by the greatest amount from that given by substitution of the basic feasible solution (when the value of the integral (3) is of course $P\times(s_{ha}-s_1)\ (z_{ka}-z_1)$) is that which makes $x_1=P-n_{ka,ha}, x_{m+ka\times ha}=0$. But this is not necessarily the step carried out by normal application of the Simplex algorithm. (The step taken is usually that which makes $x_{m+ka\times ha}=0$, but some other of the variables x_u $u=1(1)m$ than x_1 non-zero.) Nevertheless the most advantageous first step may be forced by entering the Simplex algorithm loop for the first time at the stage immediately after e has been chosen, with e artificially made to be 1.

For general values of z and s the linear function (1) will of course be non-integral, and the additional computation $[n(z,s)+1]$ must be performed.

Numerical Example

An example is provided by the continued fraction expansion

$$\log(1+z)=\frac{z}{1+}\,\frac{1^2z}{2+}\,\frac{1^2z}{3+}\,\frac{2^2z}{4+}\,\frac{2^2z}{5+}\cdots\frac{r^2z}{2r+}\,\frac{r^2z}{2r+1+}\cdots. \tag{7}$$

A specimen having the form of Table I for this expansion is given in Table II ($n_{r,r'}$ here is the order of the convergent C_n of (7))

TABLE II

z'	s	1	2	3	4	5	6	7	8	9	10
1		2	2	3	4	5	6	6	7	8	9
2		2	3	4	5	6	7	8	9	10	11
3		2	4	5	6	8	9	10	12	13	14
4		3	5	7	8	10	12	14	15	17	19
5		4	6	9	11	14	16	18	21	23	26
6		6	9	12	16	19	22	25	29	32	35

Table II shows the values of $[n(z_r', s_{r'})+1]$ $r=1(1)6$, $r'=1(1)10$ where $z=2^{z'-3}$, when the linear function (1) is the plane

$$n(z', s)=b_0+b_1z'+b_2s \tag{8}$$

and Table IV corresponding values when the linear function (1) is the quadric

$$n(z', s)=b_0+b_1z'+b_2s+b_3z'^2+b_4z's+b_5s^2 \tag{9}$$

TABLE III

z'	s	1	2	3	4	5	6	7	8	9	10
1		2	5	8	11	14	17	20	23	26	29
2		4	7	10	13	16	19	22	25	28	31
3		5	8	11	14	17	20	23	26	29	32
4		6	9	12	15	18	21	24	27	30	33
5		7	10	13	16	19	22	25	28	31	34
6		8	11	14	17	20	23	26	29	32	35

TABLE IV

z'	s	1	2	3	4	5	6	7	8	9	10
1		6	6	6	7	7	8	8	8	9	9
2		4	5	5	6	7	8	9	10	11	12
3		3	4	6	7	8	10	11	13	14	16
4		4	6	7	9	11	13	15	17	19	21
5		6	8	11	13	15	18	20	23	25	28
6		9	12	15	18	21	24	27	29	32	36

In the event the coefficients b_u in (8) and (9) determined as the solutions of linear programming problems are, for (8)

$$b_0=-2.2,\ b_1=+1.2,\ b_3=+3.0 \tag{10}$$

and for (9)

$$\begin{cases} b_0=+9.092025,\ b_1=-4.577710,\ b_2=-0.284254, \\ b_3=+0.682004,\ b_4=+0.501022,\ b_5=+0.016360 \end{cases} \tag{11}$$

This example has only been chosen to illustrate the feasibility of the method, and not for application. The logarithm of real argument should

of course be computed by TSCHEBYSCHEFF expansions as given in [4], and by means of optimal rational approximations when these have been discovered.

In conclusion there are two dangers inherent in the method which should be pointed out.

The first is that although the inequalities (2) must be satisfied, inequalities of a similar kind at points in the range $s_1 \leqslant s \leqslant s_{ha}$, $z_1 \leqslant z \leqslant z_{ka}$, other than those which are tabulated, should also hold. In the final solution of the linear programming problem some of the residual variables will be zero. In the neighbourhood of these points such inequalities may not hold. This can only be verified by judicious experimenting, repeating the solution of the linear programming problem (taking into account the new information obtained) should the results be unsatisfactory.

Secondly it must be remarked that the linear programming problem proposed is extremely ill-conditioned, and it may well occur that rounding off errors propagate to such an extent that after a few steps it is senseless to prolong the computation. This difficulty may be overcome in two ways. Firstly the initial data is available to infinite accuracy so that repetition of the computation with suitable precision will always guarantee that the mathematical and the computational realities conform. Secondly the linear function (4) and the inequalities (2) may be evaluated as a check at each stage. If (4) ceases to decrease or (2) cease to obtain (within a certain accuracy) then the intermediate set of constants b_u may be accepted as a solution. It will not of course be the optimal solution but it will be a reasonable one.

It will be noted that considerable freedom is left in the choice of the functions $\phi_u(z, s)$ occurring in (1). The convergence theory of power series and continued fractions may provide powerful hints as to which functions to choose. Numerical experience indicates that for many power series and continued fractions the rate of convergence as z varies in the complex plane is dependent upon $r = |z|$ and substantially independent of $\theta = arg(z)$. This would encourage the adoption of polar coordinates, since the profile function might very accurately be approximated by a short double series in r and θ, but not by such a series in $x = \mathrm{Re}(z)$ and $y = \mathrm{Im}(z)$.

Again, the convergence theory of continued fractions indicates that certain continued fractions converge when the argument lies in a parabolic domain in the complex plane. This implies that the equation

$$n(x, y, s_{\mathrm{const}}) = \mathrm{constant} \tag{12}$$

is approximately that of a parabola, and would encourage the choice of a system of parabolic cylinder coordinates in (1). Light will no doubt be thrown upon these speculations by subsequent work upon the computation of functions of a complex argument.

Acknowledgements

This paper was written when the author was a member of the Institute of Applied Mathematics at the University of Mainz (he is grateful to the Deutsche Forschungsgemeinschaft for a grant which enabled him to write it) and revised in Amsterdam. Provisional computations were carried out on the Z–22 in Mainz, and Tables II–IV were computed using the X1 in Amsterdam.

REFERENCES

1. Wynn, P., The Numerical Efficiency of Certain Continued Fraction Expansions, Proc. Kon. Ned. Akad. v. Wetensch., Amsterdam Series **A 65**, 127 (1962).
2. Charnes, A., W. W. Cooper and A. Henderson, Introduction to Linear Programming, Wiley, New York 1953.
3. Vajda, S., The Theory of Games and Linear Programming, Methuen, London, 1956.
4. Clenshaw, C. W., Polynomial Approximations to Elementary Functions, MTAC, VIII, 47, 143 (July 1954).