Numerische Mathematik 6, 89—97 (1964)

Zur Berechnung des Fehlerintegrals

Von
J. PATRY und J. KELLER

Zusammenfassung

Ein neuer Algorithmus zur Berechnung des Fehlerintegrals auf der reellen Achse wird erläutert. Sein Vorteil liegt darin, daß die gleiche Berechnungsformel auf der gesamten reellen Achse eine gute relative und absolute Genauigkeit ergibt. Für komplexe Werte der Unabhängigen muß man sich teilweise auf eine Reihenentwicklung, teilweise auf einen Kettenbruch stützen, dessen Konvergenz bei richtiger Wahl des letzten Gliedes wesentlich verbessert werden kann.

1. Einleitung

Das Fehlerintegral spielt in vielen Problemen (u.a. bei der Berechnung von Neutronenspektren in der Reaktorphysik) eine wesentliche Rolle. Zwei Methoden sind seit einiger Zeit bekannt, um diese Funktion mit Hilfe von Rechenautomaten näherungsweise zu berechnen [*1*]. Sie haben jedoch beide den Nachteil, daß die Funktion

$$\begin{aligned} Erfc(x) &= 1 - \Phi(x) \\ &= \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-y^2} dy \end{aligned} \tag{1}$$

nicht überall mit einer für die Anwendung genügenden Genauigkeit bestimmt werden kann.

Die Funktion $Erfc(x)$ steht in vielen Fällen nicht allein. Sie wird oft mit einem Faktor multipliziert, der so groß sein kann, daß das Produkt die Größenordnung 1 erreicht, trotzdem x größer als 4 ist. Eine gute relative Genauigkeit ist damit erforderlich. Deshalb wurden kürzlich zwei neue Methoden bekannt [*2*], die die Genauigkeitsbedingung erfüllen. Sie haben jedoch den Nachteil, daß eine Potenzreihe im Bereich

$$0 \leqq x < a \qquad (x \text{ rein reell})$$

und ein Kettenbruch im Bereich

$$a \leqq x < \infty \qquad (x \text{ rein reell})$$

$$1 \leqq a \leqq 2$$

benützt wird.

Die hier behandelte Methode stützt sich auf eine Umformung der Grundgleichung. Dadurch wird eine neue Funktion $F(x)$ definiert, die beschränkt bleibt und etwa wie (l/x) bei zunehmendem x gegen Null strebt. Sie wird durch eine Differentialgleichung definiert und kann mit Hilfe eines Kettenbruches berechnet

werden. Der letztere weist eine andere Form auf als derjenige, welcher für die unvollständige Gamma-Funktion aufgestellt wurde [3].

Der Kettenbruch hat den Vorteil, daß er analytisch bestimmt wird und damit für jeden Wert von x gültig ist, solange die Konvergenz gewährleistet ist. Der Nachteil liegt darin, daß die Konvergenz in der Nähe des Nullpunktes mangelhaft ist, so daß viele Glieder zu berücksichtigen sind.

Es hat sich gezeigt, daß eine numerisch-analytische Berechnungsmethode bei reellen x-Werten zu einem Algorithmus (*fehlin*) führt, der eine sehr gute relative Genauigkeit (besser als 10^{-7} bis $x=5$, besser als 10^{-6} für größere Werte) aufweist. Der maximale absolute Fehler beträgt $25 \cdot 10^{-9}$.

Die Konvergenz des Kettenbruches wird schließlich näher untersucht und mit Hilfe einer Extrapolationsmethode beschleunigt, so daß die rein analytische Methode, insbesondere bei komplexen x-Werten, auch mit einem vernünftigen Aufwand angewendet werden kann.

2. Die Umformung der Funktion *Erfc* (x)

Die in (1) definierte Funktion $Erfc(x)$ kann so umgeformt werden, daß einer der beiden Faktoren sich nur wenig ändert, der andere durch eine einfache analytische Form dargestellt wird:

$$\begin{aligned} Erfc(x) &= \frac{2e^{-x^2}}{\sqrt{\pi}} \int_0^\infty e^{-(2xz+z^2)}\, dz \\ &= e^{-x^2} F(x). \end{aligned} \tag{2}$$

Es wird für $F(x)$ eine Differentialgleichung und anschließend ein Kettenbruch aufgestellt.

Satz 1. Die Funktion $F(x)$ ist eine Lösung der Differentialgleichung

$$\frac{dF}{dx} - 2xF = -\frac{2}{\sqrt{\pi}} \tag{3}$$

mit

$$F(0) = 1.$$

Beweis. Aus (2) folgen die beiden Beziehungen:

$$2\int_0^\infty (x+z)\, e^{-(2xz+z^2)}\, dz = \int_0^\infty e^{-(2xz+z^2)}\, d(2xz+z^2) = 1 \tag{4}$$

und

$$\frac{dF}{dx} = -\frac{4}{\sqrt{\pi}} \int_0^\infty z\, e^{-(2xz+z^2)}\, dz. \tag{5}$$

Aus diesen Gleichungen erhält man unmittelbar die angegebene Differentialgleichung (3). Die Anfangsbedingung erfolgt aus (2).

Eine Auswertung dieser Funktion mit Hilfe von Potenzreihen erweist sich als nicht zweckmäßig, weil die Gliederzahl für eine vorbestimmte Genauigkeit bei steigendem x sehr rasch zunimmt. Bei reellen Werten von x hat die Potenzreihe noch den Nachteil, daß die Glieder alternativ positiv und negativ sind. Die Genauigkeit wird auch deshalb in vielen Fällen mangelhaft.

3. Ein Kettenbruch für $F(x)$

Der Zusammenhang zwischen $Erfc(x)$ und der vollständigen Gamma-Funktion [3]

$$Erfc(x) = \tfrac{1}{2}\Gamma(\tfrac{1}{2}, x^2)$$

führt zu einem Kettenbruch für $F(x)$

$$F(x) = \cfrac{x}{x^2+}\ \cfrac{\frac{1}{2}}{1+}\ \cfrac{1}{x^2+}\ \cfrac{\frac{3}{2}}{1+}\cdots . \tag{6}$$

Dieser Kettenbruch weist den großen Nachteil auf, daß die Glieder nicht alle die gleiche Form haben. Das ist bei Konvergenz-Betrachtungen ein großes Hindernis. Es gibt jedoch einen anderen Kettenbruch, der diesen Nachteil nicht aufweist.

Satz 2. Die Funktion $F(x)$ kann mit Hilfe des folgenden Kettenbruches definiert werden

$$F(x) = \cfrac{1}{a_0 x+}\ \cfrac{1}{a_1 x+}\ \cfrac{1}{a_2 x+}\cdots . \tag{7}$$

wobei die a_i eine monoton abnehmende Reihe positiver Zahlen bildet.

Beweis. Es werden neue Hilfsfunktionen definiert:

$$\begin{aligned} F_0(x) &= 1/F(x) \\ F_n(x) &= a_n x + 1/F_{n+1}(x). \end{aligned} \tag{8}$$

Aus (3) folgt

$$\begin{aligned} F'(x) &= \frac{F_0'}{F_0^2} = \frac{2x}{F_0} - \frac{2}{\sqrt{\pi}} \\ F_0'(x) + 2x F_0(x) &= \frac{2}{\sqrt{\pi}} F_0^2(x) . \end{aligned} \tag{9}$$

Eine allgemeine Form dieser Gleichung lautet

$$F_n'(x) + 2b_n x F_n(x) = c_n F_n^2(x) - d_n . \tag{10}$$

Aus (8) und (10) erhält man

$$F_{n+1}'(x) + 2(a_n c_n - b_n) x F_n(x) = (a_n + d_n) F_{n+1}^2(x) - c_n \tag{11}$$

mit der Bedingung

$$a_n c_n = 2b_n . \tag{12}$$

Daraus folgen die Rekursionsformeln

$$\begin{aligned} b_{n+1} &= b_n = 1 \\ c_{n+1} &= a_n + d_n \\ d_{n+1} &= c_n . \end{aligned} \tag{13}$$

Schließlich wird

$$c_{n+1} = \frac{2}{c_n} + c_{n-1} \tag{14}$$

$$a_n = \frac{2}{c_n} \tag{15}$$

mit den Anfangswerten

$$c_0 = \frac{2}{\sqrt{\pi}}$$
$$c_1 = \frac{2}{c_0} + o = \sqrt{\pi}\,. \tag{16}$$

Damit lassen sich alle a_n bestimmen.

4. Eine allgemeine Definition der a_n

Die a_n können analytisch definiert werden und stehen in engem Zusammenhang mit einer Reihe bestimmter Integrale.

Satz 3. Die a_n sind folgendermaßen definiert:

$$a_{2n} = \sqrt{\pi}\,\frac{(2n-1)!!}{(2n)!!}$$
$$a_{2n+1} = \frac{2}{\sqrt{\pi}}\,\frac{(2n)!!}{(2n+1)!!}\,, \tag{17}$$

$$a_n = \frac{1}{\sqrt{\pi}} \int_{-\pi/2}^{\pi/2} \cos^n x\,dx = 2/n \cdot a_{n-1}. \tag{18}$$

Beweis. Die Beziehungen (18) folgen unmittelbar aus (17), wenn man die Rekursionsformel (19) für die Integrale berücksichtigt:

$$\int_{-\pi/2}^{+\pi/2} \cos^n x\,dx = \frac{n-1}{n} \int_{-\pi/2}^{+\pi/2} \cos^{n-2} x\,dx. \tag{19}$$

Weiter gilt aus (15) und (17)

$$c_{2n} = \frac{2}{\sqrt{\pi}}\,\frac{(2n)!!}{(2n-1)!!}$$
$$c_{2n+1} = \sqrt{\pi}\,\frac{(2n+1)!!}{(2n)!!}\,. \tag{20}$$

Die Definitionen (17) entsprechen den Anfangswerten (15), da die Beziehungen gelten:

$$n!! = 1 \quad \text{für} \quad n = 0$$
$$n!! = 1 \quad \text{für} \quad n = \pm 1 \tag{21}$$

und

$$n!! = n \cdot (n-2)!!. \tag{22}$$

Weiter kann man feststellen, daß die Definitionen (20) die Rekursionsformel (14) genügen.

5. Die numerisch-analytische Lösung

Der Kettenbruch (7) weist den großen Nachteil auf, daß die Konvergenz in der Nähe des Nullpunktes sehr schlecht wird. Es wurde deshalb eine numerisch-analytische Methode angewandt. Die Funktion $F_1(x)$ wurde auf Grund von zwei Tabellenwerken [4] und mit Hilfe eines Ausgleichsprogrammes durch eine rationale Funktion angenähert.

$$F = \frac{1}{\sqrt{\pi}}\left[\frac{1}{x + \dfrac{1}{(2-Q)\,x + \sqrt{\pi}}}\right]. \tag{23}$$

Dadurch konnte die Funktion $F(x)$ durch eine Näherungsfunktion dargestellt werden. Es soll jedoch bemerkt werden, daß diese Näherung nur für die reelle Achse aufgestellt wurde. Bei komplexen Werten von x ist deshalb äußerste Vorsicht am Platz.

Die Beziehung wurde als ALGOL-*procedure* umgearbeitet, die zur Prüfung derselben, sowie bei anderen Programmen dienen kann:

```
real procedure fehlin(x);
real x;
begin
real t, q, z;
z := abs(x);
q := (8.5840 7657₁₀ − 1 + z × (3.0781 8193₁₀ − 1 +
     z × (6.3832 3891₁₀ − 2 − z × 1.8240 5075₁₀ − 4)))
     /(1 + z × (6.5097 4265₁₀ − 1 + z × (2.2948 4819₁₀ − 1
     + z × 3.4030 1823₁₀ − 2)));
t := 1.7724 5385 + z × (2 − q);
fehlin := exp(− z × z) × t/(1.7724 5385 × (z × t + 1));
end;
```

Dabei gilt

$$fehlin(x) \equiv Erfc(x). \tag{24}$$

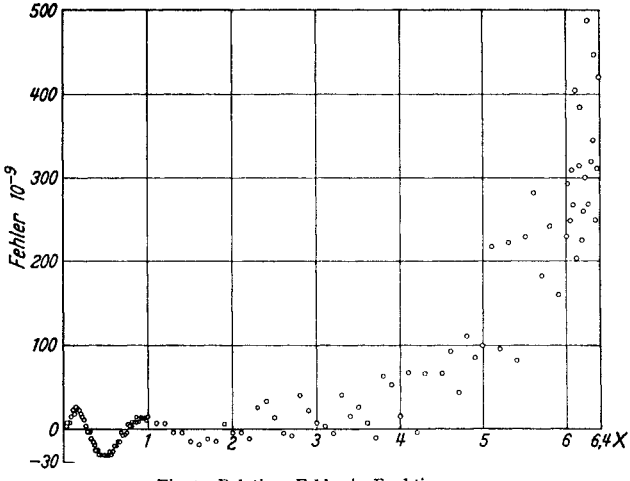

Fig. 1. Relativer Fehler in Funktion von x

Eine umfangreiche numerische Kontrolle mit den Werten (0/0.02/1.0), (1.0/0.1/6.0) und (6.0/0.02/6.38) ergab die folgenden Fehler: (Siehe Fig. 1.)

Absoluter Fehler $< 25 \times 10^{-9}$;

Relativer Fehler $< 30 \times 10^{-9}$, $x \leqq 2$;

$< 100 \times 10^{-9}$, $x \leqq 4.6$;

$< 500 \times 10^{-9}$, $x \leqq 6.38$.

Die Fehler im oberen Bereich ($x>5$) zeigen große Schwankungen. Die Vermutung liegt deshalb nahe, daß die Rundungsfehler in der Rechenanlage (Mantisse mit 29 Bits) eine große Rolle spielen. Auf alle Fälle ist die Genauigkeit wesentlich höher als bei anderen vorgeschlagenen Methoden [*1*].

6. Eine analytische Untersuchung des Kettenbruches

Der Algorithmus *fehlin* wurde nur auf der reellen Achse ausgeprüft. Wenn der Imaginärteil von x klein bleibt, ist noch eine gute Genauigkeit zu erwarten. Es wäre jedoch vorteilhaft, den Kettenbruch unmittelbar benützen zu können. Dafür soll die Konvergenz im erweiterten Sinn untersucht werden.

Es werden folgende Funktionen definiert:

$$\begin{aligned} N_n(x) &= \frac{1}{F_n(x)} = \frac{1}{a_n x + N_{n+1}(x)} \\ N_0(x) &= F(x). \end{aligned} \tag{25}$$

Dieser Kettenbruch soll so umgeformt werden, daß man die allgemeine Beziehung erhält:

$$\alpha_n = \frac{\varphi_n}{1-\alpha_{n+1}} \tag{26}$$

die eingehend untersucht wurde [*5*]. Mit

$$\alpha_n = -N_n(x)/a_{n-1}x \tag{27}$$

erhält man

$$\begin{aligned} N_n(x) &= \frac{1/a_n x}{1+N_{n+1}(x)/a_n x} \\ \varphi_n &= -\frac{1}{a_n a_{n-1} x^2} = -\frac{n}{2x^2}. \end{aligned} \tag{28}$$

Aus [*5*], Gl. 15.8 folgt

$$\begin{aligned} \varrho_1(N) &= \frac{1}{2}\left[1 - \sqrt{1+\frac{2N}{x^2}}\right] \\ \varrho_2(N) &= \frac{1}{2}\left[1 + \sqrt{1+\frac{2N}{x^2}}\right]. \end{aligned} \tag{29}$$

Die Sätze 13 und 15 aus [*5*], (S. 40 und 44) zeigen, daß der Ausdruck $\varrho_1(N)$ eine gute Näherung für α_n ergibt, und daß der relative Fehler $\Delta\alpha_n$ folgendermaßen abgeschätzt werden kann:

$$\begin{aligned} \Delta\alpha_n &= \left|\frac{\alpha_n(\text{Rechnung}) - \alpha_n(\text{Wirklichkeit})}{\alpha_n(\text{Rechnung})}\right| \\ &< \Delta\alpha_{n+1}\left|\frac{\alpha_n}{1-\alpha_n}\right| \simeq \Delta\alpha_{n+1}\left|\frac{\varrho_1(n)}{\varrho_2(n)}\right|. \end{aligned} \tag{30}$$

Der relative Fehler auf α_n nimmt deshalb bei abnehmendem n rapid ab, wenn x groß genug ist. Bei kleinen Werten von x ist diese Abnahme sehr schwach:

Für

$$\begin{aligned} \frac{2n}{x^2} &= 4 \qquad (x=1,\ n=2) \\ \frac{\Delta\alpha_n}{\Delta\alpha_{n+1}} &< \frac{\sqrt{5}-1}{\sqrt{5}+1} = 0.38. \end{aligned} \tag{31}$$

Für

$$\frac{2n}{x^2} = 40 \qquad (x = 0.5,\ n = 5)$$

$$\frac{\Delta\alpha_n}{\Delta\alpha_{n+1}} < \frac{\sqrt{41}-1}{\sqrt{41}+1} = 0.73\,.$$

Für

$$\frac{2n}{x^2} = 400 \qquad (x = 0.2,\ n = 8)$$

$$\frac{\Delta\alpha_n}{\Delta\alpha_{n+1}} < \frac{\sqrt{401}-1}{\sqrt{401}+1} = 0.905\,.$$

Das Problem wird jedoch bedeutend schwieriger, wenn x rein imaginär wird, da $\varrho_1(N)$ und $\varrho_2(N)$ in diesem Falle den gleichen Absolutbetrag aufweisen, wenn $(2N/x^2)$ groß genug wird. Die erweiterte Theorie der Kettenbrüche [6] kann hier nicht angewendet werden, weil sie sich auf die Bedingung

$$\lim_{N\to\infty} \varrho_1(N) = \varrho_0 < M \tag{32}$$

stützt. Diese Bedingung ist in unserem Falle nicht erfüllt. Die Konvergenz des Kettenbruchs konnte in diesem Bericht nicht erbracht werden.

7. Die numerische Untersuchung des Kettenbruchs

Die Ergebnisse der analytischen Untersuchung wurden auch numerisch überprüft. Ein endlicher Kettenbruch wurde berechnet, wobei sich die Zahl der Brüche zwischen 1 und 30 schrittweise vergrößerte. Der Wert von $N_N(x)$ läßt sich aus (27) und (29) ausrechnen:

$$N_N(x) = -a_{N-1}\,x\,\varrho_1(N) = \frac{2\,(a_{N-1}/a_n)^{\frac{1}{2}}}{b_N\,x + \sqrt{4+(b_N\,x)^2}} \tag{33}$$

$$b_N = \sqrt{a_N \cdot a_{N-1}}\,.$$

Da der Zähler von (33) besonders bei großen N-Werten praktisch gleich 2 wird, wurde er grundsätzlich gleich 2 gesetzt. Diese Substitution hat noch den Vorteil, daß der Wert von $N_N(x)$ bei $x=0$ gleich 1 wird, was dem Idealwert entspricht. Es zeigt sich, daß die Ergebnisse für $N=M$ und für $N=M+1$ den Endwert eingabeln, d.h. ein Wert ist zu groß, der andere zu klein. Der Unterschied gibt den größtmöglichen Fehler an. Einige Rechenergebnisse sind in Fig. 2 für b_N

Tabelle. *Werte von* a_n

n	a_n	n	a_n
0	1.7724 53851	15	0.3591 17410
	1.1283 79167		0.3480 75577
	0.8862 26926		0.3379 92857
	0.7522 52778		0.3287 38045
	0.6646 70194		0.3202 03759
5	0.6018 02223	20	0.3123 01143
	0.5538 91828		0.3049 55961
	0.5158 30477		0.2981 05637
	0.4846 55349		0.2916 97006
	0.4585 15980		0.2856 84569
10	0.4361 89814	25	0.2800 29126
	0.4168 32709		0.2746 96701
	0.3998 40663		0.2696 57677
	0.3847 68654		0.2648 86104
	0.3712 80616		0.2603 59136
		30	0.2560 56568

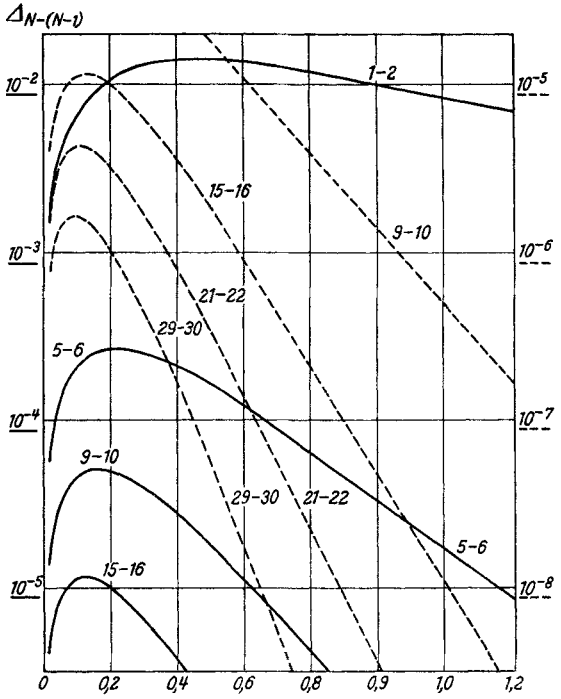

Fig. 2. $\Delta_{N-(N-1)}$ nach Gl. (35), $b_N = a_N \cdot a_{N-1}$

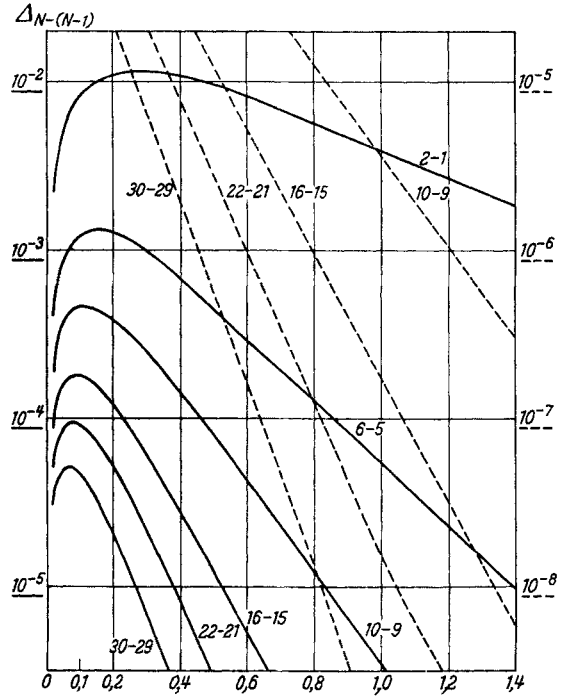

Fig. 3. $\Delta_{N-(N-1)}$ nach Gl. (35), $b_N = a_N$

aus (33) und in Fig. 3 für

$$b_N = a_N \tag{34}$$

mit

$$\Delta_{N-(N-1)} = (F(x))_N - (F(x))_{N-1} \tag{35}$$

dargestellt.

Die Kurven zeigen eindeutig die schwache Konvergenz für kleine x-Werte, die starke Konvergenz bei großen Werten. Die Kurven der Fig. 2 liegen auch, im allgemeinen, tiefer als diejenigen der Fig. 3, was der Theorie entspricht.

8. Schlußbemerkungen

Der Algorithmus *fehlin* weist bei rein reellen Werten alle Vorteile der Kettenbruch-Entwicklung auf. Es führt weiter zu den richtigen Ergebnissen auch in der Nähe des Nullpunktes, wo der Kettenbruch sehr schlecht konvergiert.

Die Berechnung des Fehlerintegrals läßt sich auch in der komplexen Ebene mit Hilfe des Kettenbruchs überall ausführen, sofern die erste Näherung $N_n(x)$ richtig ausgewählt wird. Die Konvergenz ist jedoch in einem noch zu bestimmenden Gebiet sehr schwach, so daß eine Reihen-Entwicklung günstiger ist.

Das Problem konnte nicht in seiner vollen Allgemeinheit gelöst werden. Diese Untersuchung soll jedoch einen Weg aufzeigen, der einen Schritt näher an dieses Ziel führt.

Literatur

[1] Hasting, C.: Approximations for Digital Computers, S. 167–169. Princeton 1955, und Zuse KG, Beschreibung zum Programm der Z 22, Nr. Z 63.

[2] Clenshaw, C. W., G. F. Miller and M. Woodger: Num. Math. **4**, H. 9, 403 (1963) (bes. 414) Algorithmus for special Functions I.

Reichel, A.: AEEW-R 118: Approximate Formule for the claculation of Effective Resonance Integrals in the Statistical Region, S. 18. Winfrith 1962.

Thacher Jr., H. C.: Com ACM **6**, 6, 315 (1963), Algorithm 181, Complementary error function — large x.

[3] Bateman, H.: Higher Transcendental Functions, Vol. 2, p. 136. New York: McGrav-Hill 1953.

[4] Lösch, F.: Siebenstellige Tafeln der elementaren Transzendenten Funktionen. Berlin-Göttingen-Heidelberg: Springer 1954. — National Bureau of Standard-Tables of the Error Function and its Derivative, Washington 1954.

[5] Patry, J.: Über die linearen Differentialgleichungen mit sinusförmigen Koeffizienten. Zürich: Juris-Verlag 1957.

[6] — Sur la résolution numérique dans les cas limites des equations différentielles linéaires à coefficients sinusoidaux. ZAMP **10**, 35 (1959).

Eidgenössisches Institut für Reaktorforschung
Würenlingen/Schweiz

(Eingegangen am 17. Oktober 1963)