ZASTOSOWANIA MATEMATYKI
APPLICATIONES MATHEMATICAE
XV, 1 (1976)

ALGORITHM 43

Z. CYLKOWSKI (Wrocław)

EVALUATION OF AN IMPROPER INTEGRAL ON A FINITE INTERVAL

1. Procedure declaration. The function *subtrap* evaluates the integral

$$\int_a^b f(x)\,dx. \tag{1}$$

The function $f(x)$ in (1) can be regular, it can also have singularities at the ends of the integration interval. In one of the ends it can be unlimited, however, it is then required that the value of this end point be equal to zero and that in its neighbourhood the values of $f(x)$ can be calculated with small relative error.

Data:

a, b — integration limits;
f — identifier of the integral function of type **real** with one parameter x of type **real**;
eps — relative calculation error of (1);
maxr — maximum admissible number of type **real**;
n — number which limits the length of calculations; during the calculations at most $n-\frac{1}{2}\big(1+(-1)^n\big)$ $(n \geqslant 3)$ values of the integral function will be used.

Additional result:

n — number not exceeding the value of n given in the data; if $n > 0$, the desired accuracy has been obtained after calculation of n function values, if $n = 0$, the desired accuracy has not been obtained, and if $n = -1$, then probably the integral (1) is not convergent.

Remark 1. The divergence is determined after at most $\frac{1}{2}\ln(maxr)$ evaluations of the integral function.

Remark 2. The nodes of the quadrature used are symmetrically distributed in the integration interval and are fairly much condensed in the neighbourhood of the integration interval end points. If $f(x)$ is limited and the value of *eps* is not too small, then the nodes are interior points of the integration interval. Otherwise, some of them usually merge together with the non-zero interval end point due to the limited accuracy of the computer. For complicated integral functions it is therefore advisable to programme the body of the function f in such a manner as to avoid unnecessary evaluation of the same function value.

2. Method used. Let us assume that the integral (1) is convergent and let us transform it using the substitution

$$x = \frac{a+b}{2} + \frac{b-a}{2}\operatorname{tgh} t,$$

as follows:

$$(2) \qquad \int_a^b f(x)\,dx = \frac{b-a}{2}\int_{-\infty}^{\infty} (\cosh^{-2}t) f\left(\frac{a+b}{2} + \frac{b-a}{2}\operatorname{tgh} t\right) dt$$

$$= \frac{b-a}{2}\int_0^{\infty} \cosh^{-2}t\big(f(a+c)+f(b-c)\big)\,dt, \qquad \text{where } c = (b-a)\frac{e^{-t}}{e^t+e^{-t}}.$$

From the first form of the transformed integral and from the Euler-Maclaurin formula (see, e.g., [1], p. 91-94) it can be deduced that the trapezoid formula will be an adequate quadrature. To use it, the unlimited integration interval of (2) has to be replaced by some interval $\langle 0, R\rangle$ which is sufficiently great for the relative truncation error not to exceed *eps*.

Let $\varphi(t)$ denote the integral function of (2), let

$$T = \int_0^{\infty} \varphi(t)\,dt$$

and let

$$E = \int_R^{\infty} |\varphi(t)|\,dt$$

be the truncation error. We have to find R such that

$$(3) \qquad E \leqslant |T|\,eps.$$

The value of T will be calculated from

$$(4) \qquad T \approx h\left(\frac{1}{2}\varphi(0) + \sum_{k=1}^{n}\varphi(kh)\right) \qquad (h > 0),$$

where n is some natural number.

```
real procedure subtrap(a,b,f,eps,maxr,n),
 value a,b,eps;
 integer n;
 real a,b,eps,maxr;
 real procedure f;
 begin
  integer k,N,ne;
  real ba,c,ch,ex,e1,h,h2,R,s,t,v,x;
  real procedure cft;
   begin
    k:=k+2;
    ex:=exp(x);
    e1:=1.0/ex;
    ch:=2.0/(ex+e1);
    c:=ba×e1×ch;
    c:=cft:=(f(a+c)+f(b-c))×ch×ch;
    c:=abs(c)
   end cft;
  ba:=.5×(b-a);
  x:=h2:=2.0;
  k:=2;
  t:=f(a+ba)+cft;
  if ba=.0
   then go to B;
  s:=abs(b-a);
  N:=entier(.5×(ln(maxr)+(if s≥1.0 then .0 else ln(s))));
  if N>n
   then N:=n;
  ne:=-1;
A:s:=c;
```

```
    x:=x+2.0;
    if k≥N
     then go to C;
    t:=t+cft;
    if c×s>abs(t×eps)×(s-c)
     then go to A;
    R:=x;
    v:=c;
    ne:=0;
    for h:=2.0,h2 while abs(t-s)>ch+ch do
     begin
      h2:=.5×h;
      s:=t+t;
      for x:=h2 step h until R do
       if k<n
        then t:=t+cft
        else go to D;
      ch:=abs(t×eps);
      if c×c≤ch×(c-v)
       then
       begin
        R:=R-h2;
        v:=c
       end c×c≤ch×(c-v)
     end h;
B:ne:=k-1;
C:s:=t;
D:n:=ne;
    subtrap:=ba×h2×s
   end subtrap
```

Let us put $R = nh$ and calculate n.

Since the form of the function $\varphi(t)$ suggests that, for t great,

$$|\varphi(t)| \sim ae^{-\mu t},$$

where a and μ are positive constants, we have

$$(5)\qquad E \sim \int_{nh}^{\infty} ae^{-\mu t}dt < h\sum_{k=n}^{\infty} ae^{-\mu kh} = h\frac{ae^{-\mu nh}}{1-e^{-\mu h}}$$

$$\sim h\frac{|\varphi(nh)|}{1-|\varphi(nh)/\varphi(nh-h)|} = h\frac{|\varphi(nh-h)\varphi(nh)|}{|\varphi(nh-h)|-|\varphi(nh)|}.$$

Formulae (3), (4) and (5) allow the calculation of n. Namely, it is possible, beginning with some $n = m$, to substitute $n = m, m+1, m+2, \ldots$ and to stop with that value which satisfies the inequality

$$|\varphi(nh-h)\varphi(nh)| \leqslant [|\varphi(nh-h)|-|\varphi(nh)|]\left|\frac{1}{2}\varphi(0)+\sum_{k=1}^{n}\varphi(kh)\right| eps.$$

In the function *subtrap* it is assumed $h = 2$ and $m = 2$. If the evaluated R is greater than

$$\min\{\tfrac{1}{2}\ln(maxr),\ \tfrac{1}{2}\ln(|b-a|maxr)\},$$

then the integral is assumed to be non-convergent. Otherwise, the trapezoid formula is used. The lengths of the subintervals are equal 2 at the beginning, and later on they are halved repeatedly until the required accuracy is obtained.

3. Certification. The control calculations have been performed on the Odra 1204 computer ($maxr = 6_{10}153$) for the integrals given below. The asymptotic approximations of the function $\varphi(t)$ give an insight in the form of the integral function of (2):

$$(6)\qquad \int_0^1 \arccos x\,dx = 1 \qquad (\varphi(t) \sim 2\pi e^{-2t}),$$

$$(7)\qquad \int_0^1 \ln^3 x\frac{dx}{1+x} = -5.6821970 \qquad (\varphi(t) \sim -32t^3e^{-2t}),$$

$$(8)\qquad \int_{-1}^{0}\frac{dx}{\big(-\ln(1+x)\big)^{0.95}} = 19.470085 \qquad (\varphi(t) \sim 4e^{-0.1t}),$$

$$(9)\qquad \int_0^1 \sin\frac{1}{\sqrt{x}}\,dx = 0.7570600 \qquad (\varphi(t) \sim 4\big(\sin(e^t)+\sin(1)\big)e^{-2t}),$$

$$(10)\qquad \int_0^1 \cos\frac{1}{x}\,dx = -0.08441095 \qquad (\varphi(t) \sim 4\big(\cos(e^{2t})+\cos(1)\big)e^{-2t}),$$

$$(11)\qquad \int_0^{1/e} \frac{e^x\,dx}{x(-\ln x)^{2.5}} = 0.8116404 \qquad \left(\varphi(t) \sim \frac{4e}{(2t+1)^{2.5}}\right).$$

In order to satisfy the requirements of Section 1 in case (8), the integral function was given the form

$$f(x) = \begin{cases} 0 & \text{for } x = -1, \\ \left(-\ln(1+x)\right)^{-0.95} & \text{for } -1 < x \leqslant -0.01, \\ \left(-x+\frac{1}{2}x^2-\frac{1}{3}x^3+\frac{1}{4}x^4\right)^{-0.95} & \text{for } -0.01 < x < 0. \end{cases}$$

The remaining integrals did not require such changes.

In every calculation there was assumed $n = 20\,000$. Table 1 contains the obtained integral values and the final values of n. The first three approximate values of (10) are essentially bad, probably the integration method did not "grasp" the oscillation of the function $\cos(1/x)$.

TABLE 1

eps	(6)		(7)		(8)	
10^{-1}	1.0026495	9	−5.6872117	13	18.056862	53
10^{-2}	.9982234	15	−5.6730965	25	19.312796	97
10^{-3}	.9998904	41	−5.6818375	33	19.454739	287
10^{-4}	.9999846	47	−5.6820532	73	19.468385	375
10^{-5}	.9999980	57	−5.6821922	89	19.469921	937
10^{-6}	.9999997	63	−5.6821961	95	19.470068	1115
10^{-7}	1.0000000	145	−5.6821968	105	19.470084	1305
10^{-8}	1.0000000	171	−5.6821970	235	19.411404	−1

eps	(9)		(10)		(11)	
10^{-1}	.7622373	15	−.21996577	15	.7892627	9
10^{-2}	.7581698	31	−.10036838	51	.8025381	33
10^{-3}	.7564605	121	−.09919080	77	.8103825	129
10^{-4}	.7570598	383	−.08442596	10857	.8115019	1137
10^{-5}	.7570599	853	−.08441439	0	.8494819	−1
10^{-6}	.7570596	1943	−.08441309	0	.8494819	−1

Reference

[1] P. J. Davis and P. Rabinowitz, *Numerical integration*, Waltham 1967.

INSTITUTE OF INFORMATICS
UNIVERSITY OF WROCŁAW
50-384 WROCŁAW

Received on 31. 10. 1974

ALGORYTM 43

Z. CYLKOWSKI (Wrocław)

OBLICZANIE CAŁKI NIEWŁAŚCIWEJ W PRZEDZIALE SKOŃCZONYM

STRESZCZENIE

Wartością funkcji *subtrap* jest całka (1). Funkcja podcałkowa może być regularna, ale także może mieć osobliwości na końcach przedziału całkowania. W jednym z końców może być nieograniczona, wtedy jednak wymaga się, aby był on zerem i aby w jego otoczeniu wartości funkcji $f(x)$ były obliczane z małym błędem względnym.

Dane:

a, b — granice całkowania;
f — nazwa funkcji podcałkowej typu **real** z jednym parametrem x typu **real**;
eps — błąd względny, z jakim należy obliczyć całkę (1);
maxr — największa dozwolona w maszynie liczba typu **real**;
n — liczba ograniczająca długość obliczeń; w obliczeniach użyje się najwyżej $n-\frac{1}{2}(1+(-1)^n)$ $(n \geqslant 3)$ wartości funkcji podcałkowej.

Wynik dodatkowy:

n — liczba nie większa od danego n; jeśli $n > 0$, to daną dokładność uzyskano po obliczeniu n wartości funkcji, jeśli $n = 0$, to nie uzyskano danej dokładności, a jeśli $n = -1$, to istnieje obawa, że całka (1) jest rozbieżna.

Uwaga 1. Rozbieżność stwierdza się po obliczeniu najwyżej $\frac{1}{2}\ln(maxr)$ wartości funkcji podcałkowej.

Uwaga 2. Węzły zastosowanej kwadratury są rozłożone w przedziale całkowania symetrycznie i w pobliżu jego końców są dość mocno zgęszczone. Jeśli $f(x)$ jest ograniczona, a wartość *eps* nie jest zbyt mała, to węzły są punktami wewnętrznymi przedziału całkowania. W przeciwnym razie niektóre z nich zazwyczaj zlewają się — wskutek ograniczonej dokładności maszyny — z niezerowym końcem przedziału. Dla skomplikowanych funkcji podcałkowych warto zatem w tym ostatnim przypadku zaprogramować treść funkcji f tak, aby uniknąć wielokrotnego obliczania tej samej wartości.

Użyta metoda polega na przekształceniu całki (1) do postaci (2). Nieskończony przedział całkowania zastępuje się pewnym przedziałem $\langle 0, R\rangle$ tak dużym, żeby błąd względny obcięcia był nie większy od *eps*. W celu wyznaczenia R korzysta się z faktu, że dla dużych t funkcja podcałkowa w całce (2) ma przeważnie postać wykładniczą $ae^{-\mu t}$, gdzie a i μ są pewnymi stałymi. Jeśli obliczone R jest większe od

$$\min\{\tfrac{1}{2}\ln(maxr),\ \tfrac{1}{2}\ln(|b-a|\,maxr)\},$$

to całkę uznaje się za rozbieżną. W przeciwnym razie korzysta się z wzoru trapezów. Długości podprzedziałów najpierw są równe 2, a następnie są wielokrotnie połowione, aż do osiągnięcia danej dokładności. Kilka przykładowych obliczeń zamieszczono w rozdz. 3.