Absolute and Relative Error Control in Composite Interpolatory Quadrature: the CIRQUE Algorithm

J.S.C. Prentice

Department of Applied Mathematics, University of Johannesburg,

P.O. Box 524, Auckland Park, 2006, South Africa

Tel: 271-1559-3145 E-mail: jprentice@uj.ac.za

Received: February 9, 2011 Accepted: March 3, 2011 doi:10.5539/jmr.v3n3p63

Abstract

We introduce the CIRQUE algorithm, for approximating definite integrals of continuous, univariate, real-valued functions, using positive-coefficient composite interpolatory quadrature. CIRQUE estimates and controls absolute and/or relative error, without the need for a prior estimate of the magnitude of the integral. The limiting effects of roundoff error are catered for, and CIRQUE is able to provide estimates of error bounds as output. Moreover, if these bounds are deemed too large, it is a simple matter to rerun CIRQUE once to obtain an acceptable bound. We have demonstrated the algorithm using the Trapezium rule, Simpson's rule and four-point Gauss-Legendre quadrature.

Keywords: Quadrature, Interpolatory, Composite, Relative error, Absolute error, Roundoff error, Error control

1. Introduction

Say we wish to find an approximation Q_f to a definite integral I_f of a continuous, univariate, real-valued function $f(x)$, such that

$$\frac{\left|I_f - Q_f\right|}{\left|I_f\right|} \leqslant \varepsilon, \tag{1}$$

where ε is a user-defined tolerance. In other words, we seek to control the *relative* error in the approximation.

Obviously, if we have a good estimate of the integral, $\widetilde{I_f} \approx I_f$, we can easily test whether or not the tolerance is satisfied by Q_f, and if it is not, then we can, possibly, refine Q_f. However, if we do not have $\widetilde{I_f}$, we cannot test the quality of Q_f. One way around this problem is to use a numerical method of high order to compute $\widetilde{I_f}$, and a method of lower order to compute Q_f; such an approach requires the assumption that $\widetilde{I_f}$ is suitably accurate without any means to verify such assumption and, furthermore, two different methods must be employed (Krommer and Ueberhuber, 1998).

So, the question is, can we design an algorithm that allows the relative error to be estimated and controlled without prior knowledge of I_f, and without using more than one numerical method? Such an algorithm is the subject of this paper. Our algorithm, designated CIRQUE (an acronym formed from the initial letters of **R**elative **E**rror **C**ontrol in **I**nterpolatory **QU**adrature), will also detect whether or not absolute error control is appropriate, instead of relative error control; we will provide a criterion for such a selection based on the magnitude of I_f. Also, our algorithm will cater for any limitations arising from the presence of roundoff error.

In the next section, we briefly discuss concepts, terminology and notation relevant to the algorithm. These include interpolatory quadrature, error bounds in composite interpolatory quadrature, error control, roundoff error, and a distinction between relative and absolute error control based on computational efficiency. In Section 3 we describe the CIRQUE algorithm, and thereafter we present a few numerical examples. We also discuss a procedure for refining the algorithm's estimates of error bounds. In Section 6, we list the sequence of operations in CIRQUE, as a summary of the algorithm.

2. Relevant Concepts, Terminology and Notation

In this section, we provide appropriate background information. Most of the concepts used in this paper are well-established "classical" ideas and are drawn from numerous books on numerical analysis and computational integration. Such a list, which constitutes our general bibliography, includes Burden and Faires (2011), Davis and Rabinowitz (1984), Engels (1980), Ghizetti and Ossiccini (1970), Isaacson and Keller (1994), Kincaid and Cheney (2002), Krommer and Ueberhuber (1998), Stroud (1974), and Stroud and Secrest (1966).

2.1 Interpolatory Quadrature

The integral of $f(x)$ on $[a,b]$ is denoted

$$I[f;a,b] \equiv \int_a^b f(x)\,dx. \tag{2}$$

The *quadrature* of $f(x)$ on $[a,b]$ is denoted

$$Q[f;a,b] \equiv \sum_{i=1}^{n} w_i f(x_i), \tag{3}$$

wherein w_i are the quadrature *weights*, and x_i are the quadrature *nodes*. Clearly, $Q[f;a,b]$ is a linear combination of values of f sampled at n discrete nodes on $[a,b]$, and is intended to be an approximation to $I[f;a,b]$. The precise nature of the weights depends on the nature of the quadrature but, generally speaking, they are dependent on the length of $[a,b]$. In this paper, we restrict our considerations to a particular class of quadrature, known as *interpolatory* quadrature. We will not provide details of this class here (the reader is referred to the bibliography given earlier for extensive discussions of interpolatory quadrature), suffice it to say that the form of the absolute approximation error in interpolatory quadrature is ideally suited to our objectives, and we discuss this feature in the next section. Hence, from here on, $Q[f;a,b]$ refers to interpolatory quadrature.

If $[a,b]$ is subdivided into N subintervals (denoted D_j), and $Q\left[f;D_j\right]$ is applied on each subinterval, then the integral of $f(x)$ on $[a,b]$ is simply the sum of these individual quadratures. We denote such *composite* interpolatory quadrature by

$$CQ[f;a,b] \equiv \sum_{j=1}^{N} Q\left[f;D_j\right] \equiv \sum_{N}\left[\sum_{i=1}^{n} w_i f(x_i)\right]_j, \tag{4}$$

where the double sum notation indicates the sum of quadrature on each of N subintervals D_j.

It is possible, and often convenient, to write $Q[f;a,b]$ in terms of a *stepsize* parameter h, as in

$$Q[f;a,b,h] \equiv h\sum_{i=1}^{n} c_i f(x_i), \tag{5}$$

where $c_i = w_i/h$. We choose to take this stepsize as the average separation of the nodes on $[a,b]$; we will say more about h in the next section. We must make mention of the fact that in $Q[f;a,b,h]$ the nodes x_i are not necessarily uniformly distributed on $[a,b]$; nor are the endpoints of $[a,b]$ necessarily nodes. If the nodes are uniformly/not uniformly spaced, $Q[f;a,b,h]$ is said to have *uniform/nonuniform* stepsize. If a is a node, and b is not a node, then $Q[f;a,b,h]$ is said to be *left-open*; a similar definition holds for *right-open* quadrature. If neither endpoint is a node, $Q[f;a,b,h]$ is termed *open*.

Lastly, if all weights w_i in $Q[f;a,b]$ are positive, then

$$\sum_{i=1}^{n} w_i = b-a, \tag{6}$$

a feature that we will exploit later.

2.2 Error Bounds in Composite Interpolatory Quadrature

The absolute error in composite interpolatory quadrature is bounded as

$$\left|I[f;a,b] - CQ[f;a,b,h]\right| \leqslant |A|(b-a)\max_{[a,b]}\left|f^{(\theta)}(x)\right| h^r, \tag{7}$$

where A, θ and r are dependent on the number of nodes n in the underlying quadrature rule $Q\left[f;D_j,h\right]$ (Isaacson and Keller, 1994). For example, the composite Simpson's rule has $A = \frac{1}{180}, \theta = 4$ and $r = 4$ (r indicates the *order* of the method). If we impose a tolerance ε on the approximation, then the inequality

$$|A|(b-a)\max_{[a,b]}\left|f^{(\theta)}(x)\right| h^r \leqslant \varepsilon \tag{8}$$

allows the stepsize h to be determined, such that h is consistent with the maximum possible error and the imposed tolerance. Of course, if the maximum error is less than the imposed tolerance, then the actual error will also satisfy the tolerance. Once h has been determined, it is a simple matter to compute $CQ[f;a,b,h]$. Indeed, h is given by

$$h = \frac{b-a}{(n\pm 1)\left\lceil \frac{b-a}{(n\pm 1)h^*}\right\rceil} \tag{9}$$

$$h^* = \left[\frac{\varepsilon}{|A|(b-a)\max\limits_{[a,b]}\left|f^{(\theta)}(x)\right|}\right]^{\frac{1}{r}}, \tag{10}$$

where $\lceil \cdots \rceil$ indicates rounding up to the nearest integer, and ± refers to $Q[f;a,b,h]$ as being open (+) or closed (−). If it is left-open or right-open, then we use n in place of $n \pm 1$. We note here that if the underlying quadrature rule has nonuniform stepsize, such as a Gaussian rule, then h given in (9) is the average separation of the nodes and endpoints on $[a,b]$. For uniform stepsize rules, such as Newton-Cotes rules, h in (9) is the separation of any two adjacent nodes, because in such a rule the nodes are uniformly spaced. Note that

$$N = \begin{cases} \left\lceil \frac{b-a}{(n\pm1)h^*} \right\rceil & \text{open } (+) \text{ or closed } (-) \text{ quadrature} \\ \left\lceil \frac{b-a}{nh^*} \right\rceil & \text{left- or right-open quadrature} \end{cases} \tag{11}$$

and each subinterval has length $(n \pm 1)h$ (open (+) or closed (−) quadrature) or nh (left- or right-open quadrature), so that $(n \pm 1)hN$ or nhN is the length $b - a$ of the interval of integration.

If we choose to impose a tolerance ε on the relative error, we must demand

$$\begin{aligned} \frac{\left|I[f;a,b] - CQ[f;a,b,h]\right|}{\left|I[f;a,b]\right|} &\leqslant \varepsilon \\ \Rightarrow \quad \left|I[f;a,b] - CQ[f;a,b,h]\right| &\leqslant \varepsilon \left|I[f;a,b]\right|, \end{aligned} \tag{12}$$

so that

$$\varepsilon \left|I[f;a,b]\right|$$

serves as an absolute tolerance, and h must be determined from

$$|A|(b-a)\max_{[a,b]}\left|f^{(\theta)}(x)\right|h^r \leqslant \varepsilon\left|I[f;a,b]\right|. \tag{13}$$

The difficulty is immediately obvious: we do not know $\left|I[f;a,b]\right|$.

2.3 Inclusion of Roundoff Error

Taking into account the presence of roundoff error, we write

$$CQ[f;a,b,h] = \sum_N \left[\sum_{i=1}^n w_i\left(1+\mu_{w_i}\right)f(x_i)\left(1+\mu_{f_i}\right)\right]_j, \tag{14}$$

where w_i and $f(x_i)$ are exact, μ_{w_i} is the roundoff error in the computed value of w_i, μ_{f_i} is the roundoff error in the computed value of $f(x_i)$, and the double sum notation has been explained previously. Obviously, it is now understood that $CQ[f;a,b,h]$ indicates an approximation computed with a finite-precision device.

Thus,

$$CQ[f;a,b,h] = \sum_N \left[\sum_{i=1}^n w_i f(x_i)\right]_j + \sum_N \left[\sum_{i=1}^n w_i f(x_i)\left(\mu_{w_i}+\mu_{f_i}+\mu_{w_i}\mu_{f_i}\right)\right]_j. \tag{15}$$

We define the second term on the RHS of (15) as the roundoff error RO in $CQ[f;a,b,h]$, and so

$$\begin{aligned} |RO| &\equiv \left|\sum_N \left[\sum_{i=1}^n w_i f(x_i)\left(\mu_{w_i}+\mu_{f_i}+\mu_{w_i}\mu_{f_i}\right)\right]_j\right| \\ &\leqslant \max_{[a,b]}|f(x)| \sum_N \left[\sum_{i=1}^n |w_i|\left(2\mu+\mu^2\right)\right]_j \\ &\approx 2\mu \max_{[a,b]}|f(x)| \sum_N \left[\sum_{i=1}^n |w_i|\right]_j, \end{aligned} \tag{16}$$

where μ is an upper bound on $\left|\mu_{w_i}\right|$ and $\left|\mu_{f_i}\right|$ (on our platform $\mu \sim 10^{-16}$), and in the last line we have ignored μ^2. Now, if we use a quadrature rule in which all weights are positive (so that their sum is equal to the length of the interval on which they are defined), we then have (Burden and Faires, 2011; Isaacson and Keller, 1994)

$$|RO| \leqslant 2\mu \max_{[a,b]}|f(x)|(b-a), \tag{17}$$

and if both $\max\limits_{[a,b]}|f(x)|$ and $b - a$ are less than or equal to unity, we have

$$|RO| \leqslant 2\mu. \tag{18}$$

This last case is mentioned because it will be relevant later.

As regards absolute error control, roundoff error is incorporated as

$$|A|(b-a)\max_{[a,b]}\left|f^{(\theta)}(x)\right|h^{r}+|RO|\leqslant\varepsilon, \tag{19}$$

which gives

$$h^{*}=\left[\frac{\varepsilon-|RO|}{|A|(b-a)\max\limits_{[a,b]}\left|f^{(\theta)}(x)\right|}\right]^{\frac{1}{r}}. \tag{20}$$

If $\varepsilon\leqslant|RO|$, h^{*} is negative, imaginary or zero, none of which is admissible. We therefore require that

$$\varepsilon>|RO|.$$

In other words, $|RO|$ sets a lower limit on the tolerance that can be imposed. For relative error control, the condition becomes

$$\varepsilon\left|I[f;a,b]\right|>|RO|\Rightarrow\varepsilon>\frac{|RO|}{\left|I[f;a,b]\right|}.$$

Once h^{*} has been determined, we determine a new stepsize h using (9), consistent with an integer value of N. This new stepsize h is the actual stepsize used in computing the quadrature.

A very important point must be made here: in (20), h^{*} clearly depends on $|RO|$ so that, in (11), $|RO|$ must be known in order to determine N. However, in (16), we see that $|RO|$ is computed by summing over N subintervals; this is impossible if N is not known. Only if we use positive-coefficient quadrature can we eliminate the need to know N in computing $|RO|$. We will emphasize this point again in Section 3.

2.4 Absolute and Relative Error Control

We give a criterion for choosing between absolute and relative error control. We have

$$\left|I[f;a,b]-CQ[f;a,b,h]\right|=Bh^{r}, \tag{21}$$

where B is an appropriate coefficient. Hence, for relative error control,

$$\begin{aligned} Bh_R^{r} &= \left|I[f;a,b]-CQ[f;a,b,h_R]\right|\leqslant\varepsilon\left|I[f;a,b]\right| \\ \Rightarrow\quad h_R &\leqslant\left[\frac{\varepsilon\left|I[f;a,b]\right|}{B}\right]^{\frac{1}{r}}, \end{aligned} \tag{22}$$

and, for absolute error control,

$$\begin{aligned} Bh_A^{r} &= \left|I[f;a,b]-CQ[f;a,b,h_A]\right|\leqslant\varepsilon \\ \Rightarrow\quad h_A &\leqslant\left[\frac{\varepsilon}{B}\right]^{\frac{1}{r}}. \end{aligned} \tag{23}$$

Now, if

$$\left|I[f;a,b]\right|>1, \tag{24}$$

then

$$h_R>h_A,$$

and if

$$\left|I[f;a,b]\right|\leqslant1, \tag{25}$$

then

$$h_A\geqslant h_R.$$

In the case $\left|I[f;a,b]\right|>1$, we use relative error control because of the larger stepsize. In the case $\left|I[f;a,b]\right|\leqslant1$, we use absolute error control, also due to the larger stepsize. Using a larger stepsize implies greater efficiency, because fewer nodes are required on $[a,b]$. In other words, our criterion for implementing absolute or relative error control is based on considerations of computational efficiency. The choice is clearly dictated by the magnitude of $I[f;a,b]$.

3. The CIRQUE Algorithm

We seek to estimate and control the error in a numerical approximation to

$$I\left[f;a_1,b_1\right] \equiv \int\limits_{a_1}^{b_1} f\left(x\right)dx.$$

We first make a change of variables

$$\begin{aligned} x &= \left(b_1-a_1\right)z+\left(a_1-a_2\left(b_1-a_1\right)\right) \\ &\equiv mz+c, \end{aligned} \tag{26}$$

where

$$z\in\left[a_2,b_2\equiv a_2+1\right]. \tag{27}$$

In other words, we transform the integral to one defined on a unit interval. Any unit interval will do, provided $f\left(mz+c\right)$ is bounded thereon.

So, we now have

$$\begin{aligned} I\left[m\widetilde{f};a_2,b_2\right] &\equiv \int\limits_{a_2}^{b_2}\widetilde{f}\left(z\right)mdz, \\ \widetilde{f}\left(z\right) &\equiv f\left(mz+c\right) \end{aligned} \tag{28}$$

as the integral we must approximate.

Next, we define

$$g\left(z\right)\equiv\frac{m\widetilde{f}\left(z\right)}{M}, \tag{29}$$

where

$$M=\max\left\{1,\max_{\left[a_2,b_2\right]}\left|m\widetilde{f}\left(z\right)\right|\right\}, \tag{30}$$

so that $M\geqslant 1$, always, and $\left|g\left(z\right)\right|\leqslant 1$. If $\max\limits_{\left[a_2,b_2\right]}\left|m\widetilde{f}\left(z\right)\right|\leqslant 1$, then $\left|g\left(z\right)\right|\leqslant 1$, so it is not necessary for M to be less than unity, since the purpose of M is merely to ensure that $\left|g\left(z\right)\right|\leqslant 1$. Note that

$$\begin{aligned} Mg\left(z\right)=m\widetilde{f}\left(z\right) \quad &\Rightarrow \quad MI\left[g;a_2,b_2\right]=I\left[m\widetilde{f};a_2,b_2\right] \\ &\Rightarrow \quad M=\frac{I\left[m\widetilde{f};a_2,b_2\right]}{I\left[g;a_2,b_2\right]}. \end{aligned} \tag{31}$$

Stated otherwise,

$$I\left[m\widetilde{f};a_2,b_2\right]=M\int\limits_{a_2}^{b_2}g\left(z\right)dz.$$

Now,

$$I\left[g;a_2,b_2\right]\leqslant\max_{\left[a_2,b_2\right]}\left|g\left(z\right)\right|\left(b_2-a_2\right)\leqslant 1, \tag{32}$$

and so absolute error control is appropriate for approximating $I\left[g;a_2,b_2\right]$.

We demand

$$\left|I\left[g;a_2,b_2\right]-CQ\left[g;a_2,b_2,h\right]\right|\leqslant\varepsilon_g, \tag{33}$$

where $CQ\left[g;a_2,b_2,h\right]$ is any *composite positive-coefficient interpolatory quadrature* of g on $\left[a_2,b_2\right]$, and ε_g is a user-defined tolerance with $\varepsilon_g>2\mu$ (as stated earlier, 2μ is the maximal roundoff error associated with composite positive-coefficient interpolatory quadrature of a function with magnitude less than or equal to unity on a unit interval). We use (9) and (20) to find the largest stepsize h consistent with this inequality; it goes without saying that here we use $\max\limits_{\left[a_2,b_2\right]}\left|g^{(\theta)}\left(z\right)\right|$, not $\max\limits_{\left[a_1,b_1\right]}\left|f^{(\theta)}\left(x\right)\right|$, in determining h. A useful relationship between these two quantities is

$$\max_{\left[a_2,b_2\right]}\left|g^{(\theta)}\left(z\right)\right|=\left(\frac{m^{\theta+1}}{M}\right)\max_{\left[a_1,b_1\right]}\left|f^{(\theta)}\left(x\right)\right|. \tag{34}$$

Now, we have

$$\begin{aligned} & \left|I\left[g;a_2,b_2\right]-CQ\left[g;a_2,b_2,h\right]\right| \leqslant \varepsilon_g \\ \Rightarrow\ & \left|MI\left[g;a_2,b_2\right]-MCQ\left[g;a_2,b_2,h\right]\right| \leqslant M\varepsilon_g \\ \Rightarrow\ & \left|I\left[m\widetilde{f};a_2,b_2\right]-CQ\left[m\widetilde{f};a_2,b_2\right]\right| \leqslant M\varepsilon_g. \end{aligned} \tag{35}$$

Hence,

$$\begin{aligned} & \left|I\left[m\widetilde{f};a_2,b_2\right]-CQ\left[m\widetilde{f};a_2,b_2\right]\right| \leqslant M\varepsilon_g \\ \Rightarrow\ & \left|I\left[m\widetilde{f};a_2,b_2\right]-CQ\left[m\widetilde{f};a_2,b_2\right]\right| \leqslant \left|\frac{I\left[m\widetilde{f};a_2,b_2\right]}{I\left[g;a_2,b_2\right]}\right|\varepsilon_g \\ \Rightarrow\ & \frac{\left|I\left[m\widetilde{f};a_2,b_2\right]-CQ\left[m\widetilde{f};a_2,b_2\right]\right|}{\left|I\left[m\widetilde{f};a_2,b_2\right]\right|} \leqslant \frac{\varepsilon_g}{\left|I\left[g;a_2,b_2\right]\right|} \\ \Rightarrow\ & \frac{\left|I\left[f;a_1,b_1\right]-CQ\left[f;a_1,b_1\right]\right|}{\left|I\left[f;a_1,b_1\right]\right|} \leqslant \frac{\varepsilon_g}{\left|I\left[g;a_2,b_2\right]\right|} \approx \frac{\varepsilon_g}{\left|CQ\left[g;a_2,b_2,h\right]\right|}. \end{aligned} \tag{36}$$

The last inequality is due to the fact that the change of variable preserves both the integral and the quadrature (Krommer and Ueberhuber, 1998; Stroud, 1974). The LHS of the last inequality is the relative error in $I\left[f;a_1,b_1\right] \approx MCQ\left[g;a_2,b_2,h\right]$, so that

$$\frac{\varepsilon_g}{\left|CQ\left[g;a_2,b_2,h\right]\right|} \tag{37}$$

is an estimated upper bound on this relative error. Moreover, this estimate is good if ε_g is small (because then the approximation $CQ\left[g;a_2,b_2,h\right]$ is reliable).

Note the following:

1. If we find that $\left|MCQ\left[g;a_2,b_2,h\right]\right| \leqslant 1$, then absolute error control is appropriate, and we can accept $MCQ\left[g;a_2,b_2,h\right]$ as the approximation, consistent with the absolute tolerance $M\varepsilon_g$. If $M\varepsilon_g$ is unacceptably large, then we simply redo the computation with a suitably smaller absolute tolerance ε_g.
2. Obviously, choosing ε_g so that $M\varepsilon_g$ is of acceptable magnitude would avoid having to redo the computation, in the event that absolute error control is appropriate. Note that we can only know if absolute error control is appropriate once $CQ\left[g;a_2,b_2,h\right]$ has been determined, and it may be that absolute error control is not appropriate - after all, $CQ\left[g;a_2,b_2,h\right]$ is not known *a priori*. Nevertheless, we must be careful to ensure that ε_g is not less than the maximal roundoff error 2μ. This might occur if M is very large, which probably means that relative error control would be relevant, anyway.
3. If we find that $\left|MCQ\left[g;a_2,b_2,h\right]\right| > 1$, then relative error control is appropriate. If the estimated bound $\frac{\varepsilon_g}{\left|CQ\left[g;a_2,b_2,h\right]\right|}$ on the relative error is too large, we simply repeat the algorithm, with an appropriately reduced value of ε_g (again, taking care that $\varepsilon_g \nleqslant 2\mu$).
4. Typically, we would choose $\varepsilon_g \gg 2\mu$, but we should also ensure that ε_g is small enough so that $CQ\left[g;a_2,b_2,h\right]$ may be considered a reliable approximation.

   We will say more about repeating CIRQUE with a modified ε_g in the next section.

4. Numerical Examples

We demonstrate CIRQUE with two numerical examples. In both examples, we use $\mu = 2^{-53} \approx 10^{-16}$. In this section, we will use CQ_g as shorthand for $CQ\left[g;a_2,b_2,h\right]$.

1. We approximate

   $$\int_{12}^{15} e^x dx = 3.106\ldots \times 10^6.$$

   Transforming to the interval [0, 1], we find

   $$M = 9.807\ldots \times 10^6.$$

In Table 1, we show results for CIRQUE with Trapezium ($A = \frac{1}{12}, \theta = 2, r = 2$), Simpson and four-point Gauss-Legendre quadrature (GL4, $A = 0.00022, \theta = 8, r = 8$), for various tolerances. In this table, $|\Delta_R|_{UB}$ is the estimated upper bound $\frac{\varepsilon_g}{|CQ_g|}$ on the relative error; $|\Delta_R|_T$ is the actual upper bound on the relative error; $|\Delta_A|_{UB}$ is the estimated upper bound $M\varepsilon_g$ on the absolute error; $|\Delta_A|_T$ is the actual absolute error; N is the number of subintervals on $[a_2, b_2]$ used for the composite quadrature; N_f is the total number of evaluations of the integrand $f(x)$ used in the composite quadrature; and N_f^{abs} is the number of evaluations of $f(x)$ that would have been needed if absolute error control was implemented on $[a_1, b_1]$, instead of relative error control.

For each ε_g, $|\Delta_R|_{UB}$ is the same (to the indicated precision) for each method; this is simply due to the fact that, in each case, CQ_g is sufficiently accurate. In all cases, $|\Delta_R|_T < |\Delta_R|_{UB}$ and $|\Delta_A|_T < |\Delta_A|_{UB}$, as expected. Also, $|\Delta_A|_T > |\Delta_R|_T$ because the integral has large magnitude. Values of N and N_f decrease with the order of the method, but increase with decreasing ε_g. The values of N_f^{abs} show just how computationally expensive absolute error control can be, particularly for the low-order Trapezium rule.

The interpretation of $|\Delta_R|_{UB}$ is that

$$\left|I[f; a_1, b_1]\right| \in \left[\left|MCQ_g\right|(1 - |\Delta_R|_{UB}), \left|MCQ_g\right|(1 + |\Delta_R|_{UB})\right], \tag{38}$$

i.e. the exact value lies in a $\left(\left|MCQ_g\right| |\Delta_R|_{UB}\right)$-neighbourhood of $\left|MCQ_g\right|$ or, equivalently, a $|\Delta_A|_{UB}$-neighbourhood of $\left|MCQ_g\right|$, whatever the value of $\left|I[f; a_1, b_1]\right|$ might be.

Now, let us say we are unimpressed with the estimate $|\Delta_R|_{UB} = 3.2 \times 10^{-8}$ obtained for Simpson's rule with $\varepsilon_g = 10^{-8}$, and we would prefer a bound of 10^{-9}. We simply repeat CIRQUE once using

$$\varepsilon_g = 10^{-8}\left(\frac{10^{-9}}{3.2 \times 10^{-8}}\right) = 3.125 \times 10^{-10}$$

as the new tolerance. So, if the estimated upper bound on the relative error is not acceptable, it is a simple matter to correct it. The same process holds for the absolute error bound, as well. This is because both $|\Delta_R|_{UB}$ and $|\Delta_A|_{UB}$ are proportional to ε_g. Generally, if we wish to modify $|\Delta_R|_{UB}$ and/or $|\Delta_A|_{UB}$ by a factor η, we must repeat CIRQUE with ε_g modified according to

$$\varepsilon_g \to \eta\varepsilon_g, \tag{39}$$

subject, of course, to the condition that this modified value of ε_g cannot be less than 2μ. This will yield estimated upper bounds of $\eta |\Delta_R|_{UB}$ and $\eta |\Delta_A|_{UB}$. In the above example, $\eta = \frac{10^{-9}}{3.2 \times 10^{-8}}$.

2. We approximate

$$\int_0^{2\pi} \sin x dx = 0$$

as an example of an integral for which a relative error cannot be computed. Transforming to the interval $[0, 1]$, we find

$$M = 2\pi.$$

Results are shown in Table 2, where symbols have the same meaning as in Table 1. The entries have the same character as those in Table 1, all exhibiting expected behaviour. We do not show $|\Delta_R|_{UB}$ or $|\Delta_R|_T$ because relative error control is meaningless in this example. The interpretation of $|\Delta_A|_{UB}$ is that

$$\left|I[f; a_1, b_1]\right| \in \left[\left|MCQ_g\right| - |\Delta_A|_{UB}, \left|MCQ_g\right| + |\Delta_A|_{UB}\right], \tag{40}$$

i.e. the exact value lies in a $|\Delta_A|_{UB}$-neighbourhood of $\left|MCQ_g\right|$, whatever $\left|I[f; a_1, b_1]\right|$ may be. The smallness of $|\Delta_A|_T$ in all cases is simply due to the high degree of antisymmetry present in the example, and the symmetrical node distribution in the three quadrature methods; if we did not know $|\Delta_A|_T$, the only indication of the accuracy of the approximation is the $|\Delta_A|_{UB}$-neighbourhood.

5. Bound Refinement

We see in Table 1 that $|\Delta_A|_T < |\Delta_A|_{UB}$, in some instances by two orders of magnitude. This implies that $|\Delta_A|_{UB}$ is not as tight an upper bound as we might like. The reason for this is that $|\Delta_A|_{UB}$ is, effectively, the upper bound consistent with the stepsize h^* in (20). However, the stepsize actually used in $CQ[g; a_2, b_2, h]$ is h in (9), in which $\frac{b_2 - a_2}{(n \pm 1)h^*}$ has been rounded up to the nearest integer. Consequently, we have $h \leqslant h^*$. This gives

$$M|\Delta|_{min} \leqslant |\Delta_A|_T \leqslant M|\Delta|_{max} \leqslant |\Delta_A|_{UB}, \tag{41}$$

where

$$\begin{aligned} |\Delta|_{min} &\equiv |A|\,(b_2-a_2)\min_{[a_2,b_2]}\left|g^{(\theta)}(z)\right| h^r, \\ |\Delta|_{max} &\equiv |A|\,(b_2-a_2)\max_{[a_2,b_2]}\left|g^{(\theta)}(z)\right| h^r, \\ |\Delta_A|_{UB} &= M\,|A|\,(b_2-a_2)\max_{[a_2,b_2]}\left|g^{(\theta)}(z)\right| (h^*)^r + 2M\mu = M\varepsilon_g. \end{aligned} \tag{42}$$

We see that $|\Delta|_{min}$ and $|\Delta|_{max}$ are quadrature approximation errors (in $CQ\,[g; a_2, b_2, h]$) for the actual stepsize h, whereas $|\Delta_A|_{UB}$ is the quadrature error (in $CQ\,[g; a_2, b_2, h]$) for h^* plus a roundoff term. Now, it is easy to compute $|\Delta|_{min}$ and $|\Delta|_{max}$, so we can present $[|\Delta|_{min}\,, |\Delta|_{max}]$ as an interval indicating the minimum and maximum absolute error achievable with the value of the stepsize h used in $CQ\,[g; a_2, b_2, h]$. The actual absolute error lies within these bounds. Multiplying this interval by M gives error bounds on the approximation to $\left|MCQ\,[g; a_2, b_2, h]\right| \approx \left|I\,[f; a_1, b_1]\right|$, and dividing this interval by $\left|CQ\,[g; a_2, b_2, h]\right|$ gives bounds on the relative error in the approximation. Since these bounds have been determined using the actual stepsize h, rather than h^*, we expect them to be tighter bounds.

For example, for $f(x) = e^x$ and $\varepsilon_g = 10^{-4}$, we find

$$\begin{aligned} M\,[|\Delta|_{min}\,, |\Delta|_{max}] &= [22, 441] \\ M\,[|\Delta|_{min}\,, |\Delta|_{max}] &= [2, 36] \end{aligned} \tag{43}$$

for Simpson and GL4 quadrature, respectively (we have rounded these numbers to nearest integer, for ease of presentation). In both cases, $|\Delta_A|_T \in M\,[|\Delta|_{min}\,, |\Delta|_{max}]$, and $M\,|\Delta|_{max} < |\Delta_A|_{UB}$. For the relative error, we find

$$\begin{aligned} \frac{[|\Delta|_{min},|\Delta|_{max}]}{|CQ[g;a_2,b_2,h]|} &= \left[7\times10^{-6}, 14\times10^{-5}\right] \\ \frac{[|\Delta|_{min},|\Delta|_{max}]}{|CQ[g;a_2,b_2,h]|} &= \left[6\times10^{-7}, 1\times10^{-5}\right] \end{aligned} \tag{44}$$

with $|\Delta_R|_T \in \frac{[|\Delta|_{min},|\Delta|_{max}]}{|CQ[g;a_2,b_2,h]|}$, and $\frac{|\Delta|_{max}}{|CQ[g;a_2,b_2,h]|} < |\Delta_R|_{UB}$. We have confirmed that this holds for all cases considered in the two examples (for $f(x) = \sin x$, $\min\left|g^{(\theta)}(x)\right| = 0$, so the very small values of $|\Delta_A|_T$ are accommodated).

To incorporate roundoff error into these refined bounds, we write

$$[|\Delta|_{min}\,, |\Delta|_{max} + 2\mu] \tag{45}$$

in place of $[|\Delta|_{min}\,, |\Delta|_{max}]$. Since 2μ is an *upper* bound on $|RO|$ for the quadrature on $[a_2, b_2]$, it could occur that $|RO| = 0$. Hence, we retain $|\Delta|_{min}$ as the lower bound in (45).

In summary, then, we can use $|\Delta|_{max}$ to find tighter upper bounds on both relative and absolute errors in CIRQUE, and these bounds can be presented instead of $|\Delta_R|_{UB}$ and $|\Delta_A|_{UB}$. Also, $|\Delta|_{min}$ provides a lower bound on the accuracy, although this is not of primary interest.

6. Summary of the CIRQUE Algorithm

It is worthwhile to summarize the CIRQUE algorithm, in the form of a list of the sequence of operations of the algorithm. CIRQUE requires $f(x)$, $f^{(\theta)}(x)$ (or $\max_{[a_1,b_1]}\left|f^{(\theta)}(x)\right|$, at least), $[a_1, b_1]$, $[a_2, b_2]$ and ε_g as input. The order θ of the derivative $f^{(\theta)}(x)$ must correlate with the underlying quadrature used by CIRQUE (e.g. for the Trapezium rule, $\theta = 2$). CIRQUE then performs the following:

1. Transforms the integral $\int_{a_1}^{b_1} f(x)\,dx$ to one on a unit interval $[a_2, b_2]$.

2. Determines M and normalizes the transformed integrand, so defining a new integrand $g(z)$.

3. Applies composite positive-coefficient interpolatory quadrature ($CQ\,[g; a_2, b_2, h]$) to approximate the integral of $g(z)$ on $[a_2, b_2]$, with appropriate absolute error control, subject to tolerance ε_g.

4. OUTPUT: $MCQ\,[g; a_2, b_2, h]$ is the numerical approximation to $\int_{a_1}^{b_1} f(x)\,dx$.

5. If $\left|MCQ\,[g; a_2, b_2, h]\right| > 1$, the maximal relative error is estimated as $\frac{\varepsilon_g}{|CQ[g;a_2,b_2,h]|}$, and the maximal absolute error as $M\varepsilon_g$.

6. If $\left|MCQ\,[g; a_2, b_2, h]\right| \leqslant 1$, the maximal absolute error is estimated as $M\varepsilon_g$. The relative error is not estimated in this case because such estimate could be unreliable, particularly if $\left|MCQ\,[g; a_2, b_2, h]\right|$ is close to zero.

7. The bounds in #5 and #6 could be replaced with the refined bounds $\frac{|\Delta|_{max}}{|CQ[g;a_2,b_2,h]|}$ and $M\,|\Delta|_{max}$.

8. If the bounds in #5 or #6 (or #7, for that matter) are unacceptably large, CIRQUE is repeated with an appropriately modified tolerance.

7. Concluding Comments

We have developed an algorithm, designated CIRQUE, that computes a numerical approximation to a definite integral, using composite positive-coefficient interpolatory quadrature. CIRQUE is able to distinguish between the need for absolute and relative error control, and to implement such error control, without *a priori* knowledge of the magnitude of the integral. The criterion for choosing between absolute and relative error control is based on computational efficiency. Moreover, CIRQUE can provide an *a posteriori* estimate of the maximum error incurred in the quadrature process and, if such bound is unacceptably large, it is easy to rerun CIRQUE so as to achieve an acceptable bound. Roundoff error present in the quadrature process has been taken into account in the error control algorithm. The requirements of the integrand are that it is real-valued, univariate, and continuous on the interval of integration, and that the maximum magnitude of its relevant higher-order derivative is known or can be found.

Future research efforts should concern the development of CIRQUE to handle multivariate integrands, and the possible use of the algorithm in the context of adaptive quadrature.

References

Burden, R.L., and Faires, J.D. (2011). *Numerical Analysis 9th ed.*, Brooks/Cole.

Davis, P.J., and Rabinowitz, P. (1984). *Methods of Numerical Integration 2nd ed.*, New York: Academic Press.

Engels, H. (1980). *Numerical Quadrature and Cubature*, New York: Academic Press.

Ghizetti, A., and Ossiccini, A. (1970). *Quadrature Formulae*, New York: Academic Press.

Isaacson, E., and Keller, H.B. (1994). *Analysis of Numerical Methods*, New York: Dover.

Kincaid, D., and Cheney, W. (2002). *Numerical Analysis: Mathematics of Scientific Computing 3rd ed.*, Pacific Grove: Brooks/Cole.

Krommer, A.R., and Ueberhuber, C. W. (1998). *Computational Integration*, Philadelphia: SIAM.

Stroud, A.H., (1974). *Numerical Quadrature and Solution of Ordinary Differential Equations*, Berlin: Springer-Verlag.

Stroud, A.H., and Secrest, D. (1966). *Gaussian Quadrature Formulas*, Englewood Cliffs: Prentice-Hall.

Table 1. Output generated by CIRQUE, applied to the integral $\int_{12}^{15} e^x dx$, for the various tolerances indicated, and using three different quadrature methods

	ε_g	$\|\Delta_R\|_{UB}$	$\|\Delta_R\|_T$	$\|\Delta_A\|_{UB}$	$\|\Delta_A\|_T$	N	N_f	N_f^{abs}
Trap	10^{-4}	3.2×10^{-4}	0.99×10^{-4}	9.8×10^{2}	3.1×10^{2}	87	88	156583
Simp	10^{-4}	3.2×10^{-4}	0.45×10^{-4}	9.8×10^{2}	1.4×10^{2}	5	11	351
GL4	10^{-4}	3.2×10^{-4}	2.74×10^{-6}	9.8×10^{2}	8.5	1	5	180
Trap	10^{-8}	3.2×10^{-8}	0.99×10^{-8}	0.098	0.031	8661	8662	15658108
Simp	10^{-8}	3.2×10^{-8}	0.99×10^{-8}	0.098	0.031	41	83	3485
GL4	10^{-8}	3.2×10^{-8}	5.4×10^{-10}	0.098	0.002	3	15	565
Trap	10^{-12}	3.2×10^{-12}	0.99×10^{-12}	9.8×10^{-6}	3.1×10^{-6}	866218	866219	1.5×10^{9}
Simp	10^{-12}	3.2×10^{-12}	0.99×10^{-12}	9.8×10^{-6}	3.1×10^{-6}	410	821	34833
GL4	10^{-12}	3.2×10^{-12}	6.36×10^{-13}	9.8×10^{-6}	1.98×10^{-6}	7	35	1775

Table 2. Output generated by CIRQUE, applied to the integral $\int_0^{2\pi} \sin x dx$, for various tolerances and methods, as indicated

	ε_g	$\lvert\Delta_A\rvert_{UB}$	$\lvert\Delta_A\rvert_T$	N	N_f
Trap	10^{-5}	6.3×10^{-5}	2×10^{-16}	574	575
Simp	10^{-5}	6.3×10^{-5}	1×10^{-16}	16	33
GL4	10^{-5}	6.3×10^{-5}	6×10^{-16}	2	10
Trap	10^{-9}	6.3×10^{-9}	8×10^{-16}	57358	57359
Simp	10^{-9}	6.3×10^{-9}	4×10^{-18}	153	307
GL4	10^{-9}	6.3×10^{-9}	5×10^{-16}	6	30
Trap	10^{-13}	6.3×10^{-13}	4.9×10^{-16}	5748516	5748517
Simp	10^{-13}	6.3×10^{-13}	2.3×10^{-16}	1527	3055
GL4	10^{-13}	6.3×10^{-13}	1×10^{-16}	19	95