

## 5.8 SPECIAL TOPICS IN NUMERICAL INTEGRATION

### 5.8.1 Romberg Integration

If we restrict our attention now to the case of the trapezoid rule, we can make much of the previous section on extrapolation methods more precise and, at the same time, develop a very accurate recursive procedure for approximating integrals. A fundamental result is the Euler-Maclaurin formula,<sup>10</sup> which we now state without proof.

**Theorem 5.6 (Euler-Maclaurin Formula)** *If  $f$  is sufficiently differentiable, then for any  $N > 0$  there exists a set of constants  $c_k, 1 \leq k \leq N + 1$ , such that, for some  $\xi \in [a, b]$ , the error in the trapezoid rule satisfies*

$$I(f) - T_n(f) = \gamma_1 h^2 + \gamma_2 h^4 + \dots + \gamma_N h^{2N} + c_{N+1}(b-a)h^{2N+2} f^{(2N+2)}(\xi), \quad (5.17)$$

where  $\gamma_k = c_k(f^{(2k-1)}(b) - f^{(2k-1)}(a))$ .

The significance of the Euler-Maclaurin formula is that it allows us to write the error in the trapezoid rule as a series of powers of the mesh spacing  $h$ , and then use this series to derive new quadrature rules that are more and more accurate. Note that the series expansion of the error can be carried out as far as  $f$  is differentiable, but has to be terminated; in general, it cannot be taken as an infinite series.

<sup>10</sup>Leonhard Euler (1707–1783) was one of the two greatest mathematicians of the post-Newton age, the other being Carl Friedrich Gauss. Euler was born in Basel, Switzerland, and educated at the University of Basel, at first with an eye towards following in his father's career as a minister. With the assistance of his tutor and mentor Johann Bernoulli, however, he was able to convince his father to let him pursue a career in mathematics. In 1727 Euler joined the St. Petersburg Academy of Sciences in Russia, where he remained until 1741, at which time he joined the Berlin Academy of Sciences at the invitation of the Prussian king, Frederick the Great. After some disputes with the monarch, Euler left Berlin in 1766 and returned to St. Petersburg.

Euler's contributions to mathematics are nearly unmatched in their breadth. He published an enormous amount of material in a wide variety of areas, including infinite series, special functions (a field of study that he practically invented), number theory, complex variables, and hydrodynamics. His name is attached to countless results in mathematics, from Euler's formula relating the trigonometric functions to complex exponentials, to the Euler-Cauchy differential equations, to Euler's formula relating the number of sides, edges, and vertices in a polyhedron. His influence on notation is still felt today, as it was Euler who introduced  $e$ ,  $\pi$ , and  $i = \sqrt{-1}$  into the literature as standard symbols, in addition to the use of  $\Sigma$  for denoting summations, and  $\cos$  and  $\sin$  for the cosine and sine of an angle. Euler's collected works, published between 1911 and 1975, encompass 72 volumes!

Euler's work on what we call the Euler-Maclaurin formula appears to have been done in about 1738, while he was in St. Petersburg. Although Maclaurin had first discovered the formula (see below), Euler's analysis of it went far beyond what the Scotsman had done.

Colin Maclaurin (1698–1746) was born and lived almost his entire life in Scotland. Educated at Glasgow University, he was professor of mathematics at Aberdeen from 1717 to 1725 and then went to Edinburgh. He worked in a number of areas of mathematics and is credited with writing one of the first textbooks based on Newton's calculus, *Treatise of Fluxions* (1742). The Maclaurin series appears in this book as a special case of Taylor's series.

The first few constants are

$$c_1 = -\frac{1}{12}, \quad c_2 = \frac{1}{720},$$

There is a general formula relating these to the  $s$ th Bernoulli number  $B_s$ . One consequence of the Euler-Maclaurin formula is that it can be shown to be extraordinarily accurate when applied to smooth functions. In this case, the derivatives at the endpoints

$$I(f) - T_n(f) = c_{N+1}(b-a)$$

where  $N$  is arbitrary, restrained only by the smoothness of  $f$ .

Note that if we look at the  $N = 1$  case, we

$$I(f) - T_1(f) = -\frac{1}{12}(f'(b) - f'(a))$$

which shows that the corrected trapezoid rule is more accurate than the standard trapezoid rule (see Section 5.2).

To construct an even more accurate quadrature rule, we can use the series expansion of the error in (5.17) for twice as many subintervals.

$$I(f) - T_{2n}(f) = \frac{1}{4}\gamma_1 h^2 + \frac{1}{16}\gamma_2 h^4 + \dots$$

If we multiply (5.19) by 4, subtract it from (5.17)

$$I(f) - \frac{4T_{2n}(f) - T_n(f)}{3} = b_2 h^4 + b_3 h^6 + \dots$$

where  $b_k = -\frac{1}{3}(1 - 4^{1-k})c_k(f^{(2k-1)}(b) - f^{(2k-1)}(a))$ .

Note what we have done here. The value of the error in the trapezoid rule with  $2n$  subintervals is  $O(h^4)$  accurate. But there is no reason to expect that it is  $O(h^4)$  accurate. In fact, it is  $O(h^6)$  accurate.

$$I(f) - R_{2n}(f) = b_2 h^4 + b_3 h^6 + \dots$$

for which we also then have

$$I(f) - R_{4n}(f) = \frac{1}{16}b_2 h^4 + \frac{1}{128}b_3 h^6 + \dots$$

so that, multiplying (5.22) now by 16, subtracting (5.23) from (5.22) yields

$$I(f) - \frac{16R_{4n}(f) - R_{2n}(f)}{15} = \frac{1}{15}b_2 h^4 + \frac{1}{15}b_3 h^6 + \dots$$

The first few constants are

$$c_1 = -\frac{1}{12}, \quad c_2 = \frac{1}{720}, \quad c_3 = -\frac{1}{30,240}.$$

There is a general formula relating these to the so-called Bernoulli numbers.

One consequence of the Euler-Maclaurin formula is that the trapezoid rule is shown to be extraordinarily accurate when applied to periodic functions over full periods. In this case, the derivatives at the endpoints will be equal, and thus (5.17) becomes

$$I(f) - T_n(f) = c_{N+1}(b-a)h^{2N+2}f^{(2N+2)}(\xi), \quad (5.18)$$

where  $N$  is arbitrary, restrained only by the smoothness of  $f$ . See Exercise 10 for an illustration of this.

Note that if we look at the  $N = 1$  case, we have

$$I(f) - T_n(f) = -\frac{1}{12}(f'(b) - f'(a))h^2 + c_2(b-a)h^4f^{(4)}(\xi),$$

which shows that the corrected trapezoid rule is  $O(h^4)$  accurate, which we saw experimentally in Section 5.2.

To construct an even more accurate quadrature rule from the error expansion, simply write (5.17) for twice as many subintervals (replace  $h$  by  $h/2$ ):

$$I(f) - T_{2n}(f) = \frac{1}{4}\gamma_1h^2 + \frac{1}{16}\gamma_2h^4 + \cdots + \frac{1}{4^N}\gamma_Nh^{2N} + O(h^{2N+2}). \quad (5.19)$$

If we multiply (5.19) by 4, subtract it from (5.17), and then solve for  $I(f)$ , we get

$$I(f) - \frac{4T_{2n}(f) - T_n(f)}{3} = b_2h^4 + b_3h^6 + \cdots + b_Nh^{2N} + O(h^{2N+2}), \quad (5.20)$$

where  $b_k = -\frac{1}{3}(1 - 4^{1-k})c_k(f^{(2k-1)}(b) - f^{(2k-1)}(a))$ .

Note what we have done here. The value  $\frac{1}{3}(4T_{2n}(f) - T_n(f))$  is nothing more than Richardson's extrapolated value (5.15) for  $T_n(f)$ , and the expression (5.20) shows that it is  $O(h^4)$  accurate. But there is no reason to stop here. We rewrite (5.20) as

$$I(f) - R_{2n}(f) = b_2h^4 + b_3h^6 + \cdots + b_Nh^{2N} + O(h^{2N+2}), \quad (5.21)$$

for which we also then have

$$I(f) - R_{4n}(f) = \frac{1}{16}b_2h^4 + \frac{1}{128}b_3h^6 + \cdots + \frac{1}{4^N}b_Nh^{2N} + O(h^{2N+2}), \quad (5.22)$$

so that, multiplying (5.22) now by 16, subtracting from (5.21), and solving for  $I(f)$  yields

$$I(f) - \frac{16R_{4n} - R_{2n}}{15} = a_3h^6 + \cdots + a_Nh^{2N} + O(h^{2N+2}),$$

where  $a_n = -\frac{1}{15}(1-4^{2n})b_n$ . Thus  $\frac{1}{15}(16R_n - R_{2n})$  can be viewed as yet another extrapolated approximation for the integral, and this one is  $O(h^6)$  accurate.

There is no reason not to keep going. Each step of the extrapolation process yields a new quadrature method that is more accurate than the preceding one. The process can be systematized to yield an algorithm known as *Romberg integration*. Continuing much further requires that we define some notation.

Let  $T_n^{(0)}(f)$  denote the trapezoid rule values; this is the first column of what will become a triangular array. We denote the second column by  $T_n^{(1)}(f)$ . These values are computed from the trapezoid values according to the Richardson extrapolation formula:

$$T_{2n}^{(1)}(f) = R_{2n}(f) = \frac{4T_{2n}^{(0)}(f) - T_n^{(0)}(f)}{3}.$$

Note that the indexing means that the second column will be one entry shorter than the first. Generally, then, each column is computed from the preceding one according to the formula

$$T_{2n}^{(j+1)}(f) = \frac{4^{j+1}T_{2n}^{(j)}(f) - T_n^{(j)}(f)}{4^{j+1} - 1}, \tag{5.23}$$

and the array looks like

$$\begin{array}{ccccccc} T_n^{(0)}(f) & & & & & & \\ T_{2n}^{(0)}(f) & T_{2n}^{(1)}(f) & & & & & \\ T_{4n}^{(0)}(f) & T_{4n}^{(1)}(f) & T_{4n}^{(2)}(f) & & & & \\ \vdots & \vdots & \vdots & \ddots & & & \\ T_{2^k n}^{(0)}(f) & T_{2^k n}^{(1)}(f) & \dots & \dots & T_{2^k n}^{(k)}(f) & & \end{array}$$

If the integrand is smooth enough, each step across a row of the array eliminates another power of  $h^2$  from the error expansion; i.e., the first column has error that is  $O(h^2)$ , the second column has error that is  $O(h^4)$ , the third column is  $O(h^6)$ , etc. Meanwhile, going down the array decreases  $h$  by a factor of 2 for each row: if we start with  $h = 1/2$ , then in the sixth row we have  $h = 1/64$ . The upshot of all this is that the diagonal elements of the Romberg array are very accurate approximations to the integral. A formal statement is the following, which we present without proof (see p. 328 of [3]).

**Theorem 5.7 (Romberg Integration)** Assume that  $f$  is sufficiently differentiable on  $[a, b]$ , and let  $\Theta_k(f)$  be the  $k$ th diagonal element of the Romberg array:  $\Theta_k(f) = T_{2^k n}^{(k)}(f)$ . If the mesh size for the initial trapezoid rule  $T_n^{(0)}$  is  $h$ , then

$$I(f) - \Theta_k(f) = O(4^{-k}h^{2k+2}).$$

In addition, at the same time that Romberg integration is producing accurate approximations, it can also be used to estimate the error, since for each entry in the

Romberg array we can compute an estimate:

$$E_{2n}^{(k)} = \frac{1}{3} (4T_{2n}^{(k)} - T_n^{(k)})$$

(This assumes, of course, that the integrand is smooth enough to stop the Romberg process when the estimate is accurate.) Note that the bulk of the computation in computing the first column—the trapezoid rule values—occurs in the subsequent columns, done by (5.23). The ultimate efficiency therefore depends on the rate at which the rule values are computed recursively, i.e., compute  $T_{2n}^{(0)}$  first, then  $T_n^{(0)}$ , then  $T_{2n}^{(1)}$ , etc. Fortunately, this is very easy, as the next result shows.

**Theorem 5.8** Let  $T_n(f)$  denote the trapezoid rule value for a given interval  $[a, b]$ , using  $n$  subintervals. Then we can compute  $T_{2n}(f)$ , the trapezoid rule value for  $2n$  subintervals, according to

$$T_{2n}(f) = \frac{1}{2}T_n(f) + \left(\frac{b-a}{2n}\right) \sum_{i=1}^n f\left(\frac{a+(2i-1)(b-a)}{2n}\right)$$

**Proof** The key step is to recognize that the error in the  $2n$  subinterval rule is

$$T_{2n}(f) - I(f) = \left(\frac{b-a}{n}\right) \left(\frac{1}{2}f(a) + \frac{1}{2}f(b) - \frac{1}{n} \int_a^b f(x) dx\right)$$

Thus, the rule for  $2n$  subintervals is

$$T_{2n}(f) = \left(\frac{b-a}{2n}\right) \left(\frac{1}{2}f(a) + \frac{1}{2}f(b) + \sum_{i=1}^{n-1} f\left(\frac{a+(2i+1)(b-a)}{2n}\right)\right)$$

Now manipulate:

$$\begin{aligned} T_{2n}(f) &= \left(\frac{b-a}{2n}\right) \left(\frac{1}{2}f(a) + \frac{1}{2}f(b) + \sum_{i=1}^{n-1} f\left(\frac{a+(2i+1)(b-a)}{2n}\right)\right) \\ &= \frac{1}{2} \left(\frac{b-a}{n}\right) \left(\frac{1}{2}f(a) + \frac{1}{2}f(b) + \sum_{i=1}^{n-1} f\left(\frac{a+(2i+1)(b-a)}{n}\right)\right) \\ &= \frac{1}{2} \left(\frac{b-a}{n}\right) \left(\frac{1}{2}f(a) + \frac{1}{2}f(b) + \sum_{i=1,3,5,\dots}^{2n-1} f\left(\frac{a+i(b-a)}{n}\right)\right) \end{aligned}$$

Romberg array we can compute an estimate of the error using Richardson extrapolation:

$$E_{2n}^{(k)} = \frac{T_n^{(k)} - T_{2n}^{(k)}}{4^k - 1}.$$

(This assumes, of course, that the integrand is sufficiently smooth.) This can be used to stop the Romberg process when the estimated error is sufficiently small.

Note that the bulk of the computational work in Romberg integration is involved in computing the first column—the trapezoid rule values—since the computation of the subsequent columns, done by (5.23), involves only a few simple operations. The ultimate efficiency therefore depends on our ability to rapidly compute the trapezoid rule values recursively, i.e., compute  $T_{2n}^{(0)}(f)$  from  $T_n^{(0)}(f)$  without wasted effort. Fortunately this is very easy, as the next result shows.

**Theorem 5.8** Let  $T_n(f)$  denote the trapezoid rule applied to a given function  $f$  over a given interval  $[a, b]$ , using  $n$  subintervals, with uniform mesh spacing  $h = (b - a)/n$ . Then we can compute  $T_{2n}(f)$ , the trapezoid rule using twice as many subintervals, according to

$$T_{2n}(f) = \frac{1}{2}T_n(f) + \left(\frac{b-a}{2n}\right) \sum_{j=1}^n f(a + (2j-1)(b-a)/(2n)). \quad (5.24)$$

**Proof** The key step is to recognize that we can write the trapezoid rule for  $n$  subintervals as

$$T_n(f) = \left(\frac{b-a}{n}\right) \left(\frac{1}{2}f(a) + \frac{1}{2}f(b) + \sum_{i=1}^{n-1} f(a + i(b-a)/n)\right).$$

Thus, the rule for  $2n$  subintervals is

$$T_{2n}(f) = \left(\frac{b-a}{2n}\right) \left(\frac{1}{2}f(a) + \frac{1}{2}f(b) + \sum_{i=1}^{2n-1} f(a + i(b-a)/(2n))\right).$$

Now manipulate:

$$\begin{aligned} T_{2n}(f) &= \left(\frac{b-a}{2n}\right) \left(\frac{1}{2}f(a) + \frac{1}{2}f(b) + \sum_{i=1}^{2n-1} f(a + i(b-a)/(2n))\right) \\ &= \frac{1}{2} \left(\frac{b-a}{n}\right) \left(\frac{1}{2}f(a) + \frac{1}{2}f(b) + \sum_{i=2,4,6,\dots}^{2n-1} f(a + i(b-a)/(2n))\right) \\ &\quad + \sum_{i=1,3,5,\dots}^{2n-1} f(a + i(b-a)/(2n)) \end{aligned}$$

1. Compute  $T_n^{(0)}(f) = T_n(f)$ ; this is the initial trapezoid rule computation but this is not required.
2. For  $k$  from 1 to  $N$ 
  - (a) Compute  $T_{2n}^{(k)}(f) = T_{2n}^{(k)}(f)$ ; this is the first entry on a new row of  $T$ .
  - (b) Extrapolate across the row: for  $j$  from 0 to  $k-1$ ,
    - i. Compute  $T_{2n}^{(j+1)}(f) = (4^{j+1}R_{2n}^{(j)}(f) - T_{2n}^{(j)}(f))/(4^{j+1} - 1)$ .

**Programming Hint:** Note that it is easy to encode this using only enough storage to hold the entire array. In fact, if we are careful, we can get away with storing only the new values as they are computed.

TABLE 5.12 Trapezoid Rule Integration of  $f(x) = (1+x^4)^{-1}$ ,  $[a, b] = [0, 1]$

$n$	$T_k(f)$
1	0.750000000000 ✓
2	0.8455882353 ✓
4	0.8576189079 ✓

The first extrapolation produces the first Romberg value:

$$\Theta_1(f) = T_2^{(1)}(f) = \frac{4T_2^{(0)}(f) - T_1^{(0)}(f)}{3} =$$

$$T_4^{(1)}(f) = \frac{4T_4^{(0)}(f) - T_2^{(0)}(f)}{3} = 0.86162282$$

$$\Theta_2(f) = T_4^{(2)}(f) = \frac{16T_4^{(1)}(f) - T_2^{(1)}(f)}{15}$$

$$E_4^{(1)}(f) = \frac{T_2^{(1)}(f) - T_4^{(1)}(f)}{15} = 0.00$$

Thus we are confident that the Romberg value  $\Theta_2(f)$  is accurate though it was produced with only five function evaluations.

### 5.8.2 Quadrature with Nonsmooth Integrands

So far we have assumed, in most of our development, that the integrand was as smooth as we needed it to be. In the case, and we need to understand the implication

$$\begin{aligned} &= \frac{1}{2} \left( \frac{b-a}{n} \right) \left( \frac{1}{2} f(a) + \frac{1}{2} f(b) + \sum_{j=1}^{n-1} f(a + 2j(b-a)/(2n)) \right) \\ &+ \sum_{j=1}^n f(a + (2j-1)(b-a)/(2n)) \\ &= \frac{1}{2} \left( \frac{b-a}{n} \right) \left[ \left[ \frac{1}{2} f(a) + \frac{1}{2} f(b) + \sum_{j=1}^{n-1} f(a + j(b-a)/n) \right] \right. \\ &\quad \left. + \sum_{j=1}^n f(a + (2j-1)(b-a)/(2n)) \right] \\ &= \frac{1}{2} T_n(f) + \left( \frac{b-a}{2n} \right) \left( \sum_{j=1}^n f(a + (2j-1)(b-a)/(2n)) \right), \end{aligned}$$

and we are done.  $\square$

The point of this result is that we can compute the entire first column of the Romberg array using the minimum number of function evaluations. A naive implementation, with  $T_{2n}^{(0)}(f)$  computed from scratch, would require that we recompute  $n$  values of  $f$  that we had already computed in finding  $T_n^{(0)}(f)$ . As a result of this theorem, we can compute  $T_{2n}^{(k)}(f)$  using no more function evaluations than for the ordinary trapezoid rule, yet we achieve much more accuracy.

Table 5.11 shows the application of Romberg integration to our standard example of integrating  $e^x$  over the interval  $[0, 1]$ . Note the extremely rapid decay of the error: We close this section with an algorithm (in outline form) for Romberg integration (see Algorithm 5.2).

### EXAMPLE 5.16

To illustrate the computation with a minimum of extraneous effort, let's refer to Table 5.12, which gives trapezoid rule values (i.e.,  $T_n^{(0)}(f)$  values) for approximating

$$I(f) = \int_0^1 \frac{dx}{1+x^4} = 0.8669729871. \quad \checkmark$$

TABLE 5.11 Romberg Integration of  $f(x) = e^x$ ,  $[a, b] = [0, 1]$

$n = 2^k$	$\Theta_k(f)$	$I(f) - \Theta_k(f)$
1	1.859140914222952	0.14085908577048E+00
2	1.71886115187659	0.57932341754729E-03
4	1.71828268792476	0.85946571215523E-06
8	1.71828182879453	0.33548497313518E-09
16	1.71828182845908	0.32640556923980E-13
32	1.71828182845905	0.44408920985006E-15
64	1.71828182845905	-0.444408920985006E-15

Algorithm 5.2 Romberg Integration

1. Compute  $T_n^{(0)}(f) = T_n(f)$ ; this is the initial trapezoid rule computation, using  $n$  subintervals. Often  $n = 1$ , but this is not required.
2. For  $k$  from 1 to  $N$ 
  - (a) Compute  $T_{2^k n}(f) = T_{2^k}^{(0)}(f)$ ; this is the first entry on a new row of the Romberg array.
  - (b) Extrapolate across the row: for  $j$  from 0 to  $k - 1$ ,
    - i. Compute  $T_{2^{k-j} n}^{(j+1)}(f) = (4^{j+1} T_{2^k}^{(j)}(f) - T_{2^{k-j} n}^{(j)}(f)) / (4^{j+1} - 1)$ .

**Programming Hint:** Note that it is easy to encode this using only enough storage for two rows in the Romberg array; we don't need to store the entire array. In fact, if we are careful, we can get away with storing only a single row of the array, and over-writing it with the new values as they are computed.

TABLE 5.12 Trapezoid Rule Integration of  $f(x) = (1 + x^4)^{-1}$ ,  $[a, b] = [0, 1]$

$n$	$T_k(f)$
1	0.75000000000 ✓
2	0.8455882353 ✓
4	0.8576188079 ✓

The first extrapolation produces the first Romberg value:

$$\Theta_1(f) = T_2^{(1)}(f) = \frac{4T_2^{(0)}(f) - T_1^{(0)}(f)}{3} = 0.8774509800. ✓$$

It takes two extrapolations to produce the second Romberg value:

$$T_4^{(1)}(f) = \frac{4T_4^{(0)}(f) - T_2^{(0)}(f)}{3} = 0.8616289989,$$

$$\Theta_2(f) = T_4^{(2)}(f) = \frac{16T_4^{(1)}(f) - T_2^{(1)}(f)}{15} = 0.8605742004. ✗$$

The Richardson error estimate for  $T_4^{(1)}(f)$  and  $T_2^{(1)}(f)$  is given by

$$E_4^{(1)}(f) = \frac{T_2^{(1)}(f) - T_4^{(1)}(f)}{15} = 0.0010547987.4$$

Thus we are confident that the Romberg value  $\Theta_2(f)$  is accurate to within about  $10^{-3}$ , even though it was produced with only five function evaluations. ■

5.8.2 Quadrature with Nonsmooth Integrands

So far we have assumed, in most of our developments, that the function being integrated (the *integrand*) was as smooth as we needed it to be. But this will not always be the case, and we need to understand the implications of this.