Math 373: 01 — Fall 2000
MW8 SC–205
Prof. Bumby

Introduction to Numerical Integration

The treatment of integration in the text lacks the emphasis to distinguish the useful methods from those with similar formulas that have serious flaws. The formulas are also described in such a way that the interpretation of the variables in the statements is sometimes different from what it appears to be. Since numerical methods are sensitive to the order in which computations are performed, it is not sufficient to have describe the process by an expression that can only be guaranteed to be correct if the quantities it contains can be known *exactly*. In addition to controlling *error*, the computation should be described in a way that guards against the *blunder* of misinterpreting a quantity in the formula. For this reason, the process of approximating integrals will be presented here in terms of some general principles rather than in terms of formulas.

Definite Integrals. In elementary calculus, most of the emphasis is on the indefinite integral and its calculation by finding a function whose derivative is the given integrand. Although there are warnings that many expressions cannot be integrated in terms of the functions that you know, you usually don't meet such integrals in textbook exercises or exam questions. You are even encouraged to signal the end of your computation of the integral by adding "$+C$" to the function that you believe to be the answer.

When it comes to applications, the integral sign is decorated with "limits of integration" and the indefinite integral must be evaluated at these limits and the lower value subtracted from the upper. One effect of this is to make your $+C$ irrelevant. If the integral requires a substitution in its evaluation, you may either express the answer in terms of the original variable before evaluation (as in the determination of the indefinite integral) or develop rules for applying the substitution to the limits of integration also. The latter approach allows the integral to be simplified by substitution. You should review the discussion of this in your Calculus text so you have some examples handy when it is used here.

In the theoretical sections of the course, which usually have little relevance to the exam problems, the definite integral is *defined* as a limit of sums, and the indefinite integral appears as a definite integral with a variable upper endpoint.

In numerical work, only definite integrals can be calculated, and the methods resemble the sums whose limit is used in the theoretical definition.

Averages. Calculus textbooks usually include a definition of the "average value of the function f on the interval $[a, b]$" as

$$\frac{1}{b-a}\int_a^b f(x)\,dx.$$

At the time it seems like just another definition that must be memorized. However, it will be extremely useful for organizing the numerical calculation of integrals. In a sense, the definition should be turned around so that the integral from a to b becomes that average on $[a, b]$ times $(b-a)$.

The usual use of the word "average" applies to a finite list of values. In this context, the average is the *sum* of the values divided by the *count* of the values. An alternate description is that each value should be multiplied by the reciprocal of the count, and these quantities added. A slightly more general notion of average, which simulates repetitions in the list, is the *weighted average* in which the given values V_i are

multiplied by weights w_i whose sum is 1 so that

$$\text{Avg.} = \sum w_i V_i \text{ where } \sum w_i = 1.$$

Frequently, we require that all $w_i \geq 0$, because it is then an easy consequence that

$$\min V_i \leq \text{avg } V_i \leq \max V_i.$$

All numerical formulas approximate the average of a function on an interval by a weighted average of a list of values of the function on that interval.

The trapezoidal rule The first idea for approximating the integral of a function on an interval $[a, b]$ is to compute the values of the function at a small number of points and integrate the interpolating polynomial taking these values. We have estimated the interpolation error, so the accuracy of this approximation can be determined by integrating the expression for the interpolation error. In these error estimates, the $f^{(k)}(\xi)$ should be replaced by its *largest* value on $[a, b]$, or even by a deliberate overestimate produced by ignoring the signs of the individual terms in an expression for this derivative and using the triangle inequality to bound the absolute value of the sum, to obtain something that is known to be an upper bound on the error. This assures that any correct use of the formula will be more accurate. This allows you to use known data to detect bugs in your program as well as to adjust parameters to assure that it will produce reliable results when used for new results. Although the first formulas obtained have significant limitations, the error estimates contain some encouraging features that will be used in developing practical methods of numerical integration.

If we use only the endpoints a and b, we are interpolating by a straight line and the interpretation of the integral as an area shows that the integral is the area of the trapezoid bounded by the interval $[a.b]$ on the x axis, the lines $x = a$ and $x = b$ and the line joining $(a,\ f(a))$ to $(b,\ f(b))$ which approximates the graph of $y = f(x)$. The corresponding expression for the average is

$$\frac{f(a) + f(b)}{2}.$$

The error in the interpolation formula is $f''(\xi)(x-a)(x-b)/2$. The fixed part of this expression, $(x-a)(x-b)$ is a function that is negative for all x between a and b (and zero at the endpoints). This means that

$$\frac{\max f''(x)}{2}(x-a)(x-b) \leq \frac{f''(\xi)}{2}(x-a)(x-b) \leq \frac{\min f''(x)}{2}(x-a)(x-b)$$

The extreme values of the function f'' are just numbers, so the integral of the interpolation error is bounded by constant multiples of the integral of $(x-a)(x-b)$. This integral is $-(b-a)^3/3$, so the error in the integral is between $-(b-a)^3/6$ times the extreme values of $f''(x)$. This will be a *value* attained by the second derivative somewhere in $[a, b]$.

To get the error estimate for the average of f, it is only necessary to divide this by $(b-a)$, so linear interpolation has an error estimate of the form $-f''(\xi)(b-a)^2/6$ for the average. If the sign of $f''(\xi)$ does not change as ξ runs from a to b, this expression allows us to predict the sign of the error as well as its size. Thus, a positive second derivative signifies that the graph is *concave upward*, which implies that all chords lie above the curve. Linear interpolation, on which the trapezoidal rule is based, uses these chords, so it is sure to be larger than the function. Although we insist on using a *worst case* analysis in creating error estimates, retaining information about the shape of the interpolation error bound gives a bound for the error

in determining the average that is one-third of the estimate of the worst error in estimating a value of the function.

The error in estimates of averages have a similar form to the error in estimating individual values of the function, but include an extra small numerical factor. They also agree in sign with the interpolation error when that has a consistent sign.

Normalized constructions. When more terms are included, the expressions for the integral of the interpolating polynomial and for the error terms get more complicated. Performing these integrals in terms of the parameters a and b leads to a large number of terms that *eventually* combine to a power of $(b-a)$. It would be nice if the terms could be obtained from the beginning in a simple form.

If x is a variable on the interval $[a, b]$, we can write

$$x = a + t(b-a)$$

where t is a variable on the interval $[0, 1]$. That is $t = 0$ corresponds to $x = a$, $t = 1$ corresponds to $x = b$, and in any integral,

$$dx = (b-a)\,dt.$$

Moreover, if our integrand is a product of factors of the form $(x - x_i)$, each factor will be $(t - t_i)$, where

$$x_i = a + t_i(b-a).$$

That is,

$$\int_a^b \prod_{i=1}^{n}(x - x_i)\,dx = (b-a)^{n+1}\int_0^1 \prod_{i=1}^{n}(t - t_i)\,dt,$$

when n factors of $(b-a)$ come from the terms of the product and one comes from the dx. When restated for averages, one lower power of $(b-a)$ occurs since the average in x needs to be divided by $(b-a)$, but the average in t is equal to the integral since the interval has length 1.

The average of a function that is a product of differences of the variable to a list of points is the product of a constant depending on the relative position of the points in the interval with the length of the interval raised to a power equal to the number of factors in the product.

Using the Lagrange form. If the interpolation formula is expressed in Lagrange form, the function values are multiplied by terms that also have factors of $x_i - x_j$ in their denominators. Normalization changes these to $(b-a)(t_i - t_j)$ and all factors of $(b-a)$ in the change-of-variables formula disappear. That is, the multiplier of $f(x_i)$ in the computation of the *average* (the expression for the integral must retain one factor of the length of the interval) can be done using the relative position of the points in a normalized interval. Since the sum of the polynomials appearing in the Lagrange form corresponds to the unique polynomial of lowest degree that interpolates a function $g(x)$ with $g(x_i) = 1$ at the given set of x_i, it must be the constant function 1 whose average is 1.

The error term behaves differently because it is expressed in term of a high order derivative of the function instead of individual values of the the function. If the interval were to be normalized and derivatives taken with respect to t instead of x, the chain rule would generate a factor of $(b-a)$ when each derivative was computed. Since these are constant factors, they remain as factors through the calculation of later derivatives. The factors of $(b-a)$ that appeared from normaling the integral of the error term can also be though of as arising from transferring the function to the normalized interval. The other factor in the error term is found by integrating a product of all $(t - t_j)$ from 0 to 1 and dividing by a suitable factorial.

Any average computed from the Lagrange interpolation formula is given by multiplying function values by some coefficients whose sum is 1. The order of the derivative in the error term is exactly matched by a factor that would be produced by differentiating with respect to the parameter on the normalized interval. The other factor depends only of the relative position of the x_i in $[a, b]$ and can be calculated on the normalized interval.

Simpson's rule. Now, interpolate a quadratic on $[0, 1]$ using the points 0, $\frac{1}{2}$ and 1. It is easy to calculate that

$$\int_0^1 \frac{(t-\frac{1}{2})(t-1)}{(0-\frac{1}{2})(0-1)}\,dt = \frac{1}{6}$$
$$\int_0^1 \frac{(t-0)(t-1)}{(\frac{1}{2}-0)(\frac{1}{2}-1)}\,dt = \frac{2}{3}$$
$$\int_0^1 \frac{(t-0)(t-\frac{1}{2})}{(1-0)(1-\frac{1}{2})}\,dt = \frac{1}{6}$$

The integrand in the error term is

$$\frac{1}{6}(t-0)(t-\tfrac{1}{2})(t-1)$$

times an expression based on the $f'''(\xi)$ factor, that it is sampling values of the third derivative on the interval. However, the part that we have displayed has integral zero, so the formula would be *exact* when applied to a polynomial of degree 4. However, our conservative strategy for estimating the error would require us to prepare for a function whose sign exactly matched the displayed factor, so we use a bound of the largest absolute value of the third derivative times

$$\frac{1}{6}\int_0^1 \left|(t-0)(t-\tfrac{1}{2})(t-1)\right|\,dt = \frac{1}{3}\int_0^{\frac{1}{2}} (t-0)(t-\tfrac{1}{2})(t-1)\,dt = \frac{1}{192}$$

This seems pretty good, as long as the third derivative is small, but we know that we can do better if the third derivative is large, but almost constant, i.e. if the *fourth* derivative is small. How can we modify the formula?

The answer is to count the point $\frac{1}{2}$ twice. Instead of computing a new Lagrange-Hermite formula in this case, divided differences can be used to identify the new term as

$$f[a, \frac{a+b}{2}, b, \frac{a+b}{2}](x-a)(x-\frac{a+b}{2})(x-b).$$

Since divided differences do not depend on the order of the quantities in the bracket, the divided difference could be computed as

$$f[a, \frac{a+b}{2}, \frac{a+b}{2}, b]$$

using

$$f[\frac{a+b}{2}, \frac{a+b}{2}] = f'\left(\frac{a+b}{2}\right)$$

when it is needed in the computation. However, after normalization, this term leads to a *constant* multiple of

$$\int_0^1 (t-0)(t-\tfrac{1}{2})(t-1)\,dt = 0,$$

so the computation of the divided difference need not be performed. Only the existence of this constant, not its value, is needed.

The error term has an integrand of $f^{(4)}(\xi)$ multiplied by four $(x - x_i)$ factors. Our strategy calls for replacing $f^{(4)}(\xi)$ by its worst value. Then, normalization produces gives a factor of $(b-a)^4$ and the integral

$$\int_0^1 \frac{-1}{24}(t-0)(t-\tfrac{1}{2})^2(t-1)\,dt = \frac{1}{2880}$$

where the negative constant in the integrand has been inserted to correct for the function being everywhere negative. The expression for the error given in the textbook differs by a factor of 32 because it is expressed in terms of $h = (b-a)/2$.

> **If the error term would be zero if a derivative appearing in it were constant, it is often possible to extend the formula to allow a new main term that is identically zero with a higher order error term.**

Higher order Newton-Cotes rules This process can be continued by interpolating at more points. However, this has limited value for three reasons. First, the interpolating polynomials don't always give better approximations to the functions. Second, the error term involves a function with frequent changes of sign, and the integral of its absolute value may be very much larger than the integral of the function itself, with no easy way to take advantage of a tendency of the $f^{(n+1)}(\xi)$ factor not to be change very much. The third reason is related to this — the expressions giving the coefficients of the function values are integrals of functions with many changes of sign and there is no reason to expect that they will all be positive. Since we *really* want the numerical method to average the function by averaging function values, this can lead to surprising values. The changing of signs first appears when 9 equally spaced points are used to interpolate a polynomial of degree 8.

> **Interpolation is a good idea, but not a great idea. Polynomials of high degree can take unexpected turns that are not completely smoothed out by integrating. The low degree approximations have better numerical properties because the error term involves expressions that do not change sign, so they average the value of the derivative appearing in them. This allows the error term to be expressed in terms of the value of the derivative somewhere in the interval.**

Exercise S8. Explore the use of the cubic Hermite interpolation to obtain an integration rule. The **Hermite polynomials** of degree 3 that multiply $f(a)$, $f(b)$, $f'(a)$, and $f'(b)$ can be found in the proof of Theorem 3.9 of the text. Alternatively, the cubic interpolating polynomial $P_3(x)$ of degree 3 can be written as

$$P_3(x) = P_1(x) + (x-a)(x-b)Q_1(x),$$

where $P_1(x)$ is the linear interpolation, and $Q_1(x)$ is another linear polynomial that can be characterized by values at a and b. (**a**) Find an expression for $P_3'(x)$ and use it to express $Q_1(a)$ and $Q_1(b)$ as divided differences. (**b**) Evaluate

$$\int_a^b P_3(x)\,dx$$

either directly, or by using the results of (a), in terms of $f(a)$, $f(b)$, $f'(a)$, and $f'(b)$. (**c**) Evaluate

$$\int_0^1 t^2(t-1)^2\,dt$$

and apply it to give an expression for the error in using the result of (b) to approximate the average of $f(x)$ on $[a, b]$.

Composite rules. How can we get both the benefits of small steps between sample points and the nice numerical properties of low degree approximations? The same way that we did it when we were using interpolation to calculate function values! Each interpolation formula will be used on a small piece of interval. That is, we build up the formula in two steps. First, we divide the interval from a to b into m small intervals but some points u_i with $u_0 = a$ and $u_m = b$. This refinement will give us *small* intervals $[u_{i-1}, u_i]$ on which to use our approximate formula.

We have indicated that we should organized the process in terms of computing an average rather than an integral. Thus

$$\begin{aligned}
\text{avg}_{[a,b]}(f) &= \frac{1}{b-a}\int_a^b f(x)\,dx \\
&= \frac{1}{b-a}\sum_{i=1}^{m}\int_{u_{i-1}}^{u_i} f(x)\,dx \\
&= \sum_{i=1}^{m}\frac{u_i - u_{i-1}}{b-a}\frac{1}{u_i - u_{i-1}}\int_{u_{i-1}}^{u_i} f(x)\,dx \\
&= \sum_{i=1}^{m}\frac{u_i - u_{i-1}}{b-a}\,\text{avg}_{[u_{i-1},u_i]}(f)
\end{aligned}$$

Thus, the average of f between a and b is the average of the averages on the $[u_{i-1}, u_i]$, weighted by their lengths. If the u_i are equally spaced, this is the ordinary average of these smaller integrals.

Now, each $[u_{i-1}, u_i]$ may be further divided by some points to apply a simple integration rule (e.g., we can insert the midpoint of the interval in order to use Simpson's rule). Traditionally, all the points at which the function is sampled were given the same description even if they acquired different weights in the formula. In this two-stage description, the strategy to be applied on the subintervals and the choice of the subintervals are considered independent. One awkward effect is that the formulas then tend to be described in terms of the choice of the u_i rather than in terms of the x_j at which the function is to be sampled. We have already seen that this gives a factor of 32 in the expression for the Simpson's rule error term in Section 4.3 (although it turns into only a factor of 16 between our expression and the one in Section 4.4).

In computation, when you are using $f(u_i)$ as the contribution from the left endpoint of interval $i+1$, you want to remember the value that you computed when it was used as the contribution from the right endpoint of interval i, since evaluation of f is likely to be the most costly part of the evaluation. If the whole process is to be expressed in a single formula, this requires collecting these terms. The resulting composite trapezoidal rule treats endpoints different from the interior points, and Simpson's rule continues this distinction of endpoints while introducing a new classification of points of odd or even index. These formulas are probably so familiar that this curious pattern is no longer mysterious. However, there may be other ways to tell your program to remember not to calculate values twice, so this consequence of the use of a single formula should not be made to seem important.

Description and analysis of integration formulas should use the two-stage process. Each quantity appearing should be described in terms of its role in the calculation. In different applications, one or the other part of the process may be fixed and only the other part needs to be chosen to achieve required accuracy. The details of efficient computation may be designed without including everything that needs to be considered in a single formula.