

6

Discrete Approximations of the Continuum

While mathematical functions can be defined on a continuous variable, any numerical representation is limited to a finite number of values. This discretization of the continuum is the source of profound issues for numerical interpolation, differentiation, and integration.

6.1 Differentiation

A function can be locally described by its Taylor expansion:

$$f(x+h) = f(x) + f'(x)h + f''(x)\frac{h^2}{2} + \dots + f^{(n)}(x)\frac{h^n}{n!} + f^{(n+1)}(x+\vartheta)\frac{h^{n+1}}{(n+1)!}$$

The very last term is evaluated at $x + \vartheta$, which lies somewhere between x and $x + h$. Since ϑ is unknown, this last term provides a bound on the error when the series is truncated after n terms. For example, $n = 0$ tells us that $|f(x+h) - f(x)| \leq Mh$, where $M = \max_{0 \leq \vartheta \leq h} |f'(x + \vartheta)|$.

A function is said to be “of order p ”, $O(h^p)$, when for sufficiently small h its absolute value is smaller than a constant times h^p . The derivative of a function can be approximated by a difference over a finite distance, $f'(x) = [f(x+h) - f(x)]/h + O(h)$, the “forward difference” formula, or $f'(x) = [f(x) - f(x-h)]/h + O(h)$, the “backward difference” formula. Another possibility is the “center difference”

$$f'(x) = \frac{f(x+h) - f(x-h)}{2h} + O(h^2).$$

At first sight it may appear awkward that the center point, $f(x)$, is absent from the difference formula. A parabola passing through two points is

not uniquely determined; it would require $f(x+h)$, $f(x)$, and $f(x-h)$. However, it is easily shown that the slope at the center of such a parabola is independent of $f(x)$. Thus, it makes sense that the center point does not appear in the finite difference formula for the first derivative. The center difference is accurate to $O(h^2)$, not just $O(h)$ as the one-sided differences are, because the f'' terms in the Taylor expansions of $f(x+h)$ and $f(x-h)$ cancel.

The second derivative can also be approximated with a finite difference formula, $f''(x) \approx c_1 f(x+h) + c_2 f(x) + c_3 f(x-h)$, where the coefficients c_1 , c_2 , and c_3 can be determined with Taylor expansions. This is a general method to derive finite difference formulas. We find

$$f''(x) = \frac{f(x-h) - 2f(x) + f(x+h)}{h^2} + O(h^2).$$

A mnemonic for this expression is the difference between one-sided first derivatives: $\{[f(x+h) - f(x)]/h - [f(x) - f(x-h)]/h\}/h$. With three coefficients, c_1 , c_2 , and c_3 , we only expect to match the first three terms in the Taylor expansions, but the next order, involving $f'''(x)$, vanishes automatically. Hence, the leading error term is $O(h^4)/h^2 = O(h^2)$.

With more points (a larger “stencil”) the accuracy of a finite-difference approximation can be increased, at least as long as the high-order derivative that enters the error bound is not outrageously large. The error term involves a higher power of h , but also a higher derivative of f .

6.2 Verifying the Convergence of a Method

Consider numerical differentiation with a simple finite-difference: $u(x) = [f(x+h) - f(x-h)]/2h$. With a Taylor expansion we can immediately verify that $u(x) = f'(x) + O(h^2)$. For small h , this formula provides therefore an approximation to the first derivative of f . When the resolution is doubled, the discretization error, $O(h^2)$, decreases by a factor of 4. Since the error decreases with the square of the interval h , the method is said to converge with “second order.” In general, when the discretization error is $O(h^p)$ then p is called the “order of convergence” of the method.

The resolution can be expressed in terms of the number of grid points N , which is simply inversely proportional to h . To verify the convergence

of a numerical approximation, the error can be defined as some overall difference between the solution at resolution $2N$ and at resolution N . Ideal would be the difference to the exact solution, but the solution at infinite resolution is usually unavailable, because otherwise we would not need numerics. “Norms” (denoted by $\|\cdot\|$) provide a general notion of the magnitude of numbers, vectors, matrices, or functions. One example of a norm is the root-mean-square $\|y\| = \sqrt{\sum_{j=1}^N (y(jh))^2/N}$. Norms of differences therefore describe the overall difference, deviation, or error. The ratio of errors, $\|u_N - u_{N/2}\|/\|u_{2N} - u_N\|$, must converge to 2^p , where p is the order of convergence. Table I shows a convergence test for the center difference formula shown above applied to an example function. The error $E(N) = \|u_{2N} - u_N\|$ becomes indeed smaller and smaller with a ratio closer and closer to 4.

N	$E(N)$	$E(N/2)/E(N)$
20	0.005289	
40	0.001292	4.09412
80	0.0003201	4.03556
160	7.978E-05	4.01257

Table 6-I: *Convergence test for the first derivative of the function $f(x) = \sin(2x - 0.17) + 0.3 \cos(3.4x + 0.1)$ in the interval 0 to 1. The error (second column) decreases with increasing resolution and the method therefore converges. Doubling the resolution reduces the error by a factor of four (third column), indicating the finite-difference expression is accurate to second order.*

The table is all that is needed to verify convergence. For deeper insight however the errors are plotted for a wider range of resolutions in figure 1. The line shown has slope -2 on a log-log plot and the convergence is overwhelming. The bend at the bottom is the roundoff limitation. Beyond this resolution the leading error is not discretization but roundoff. If the resolution is increased further, the result becomes less accurate. For a method with high-order convergence this roundoff limitation may be reached already at modest resolution. A calculation at low resolution can hence be more accurate than a calculation at high resolution!

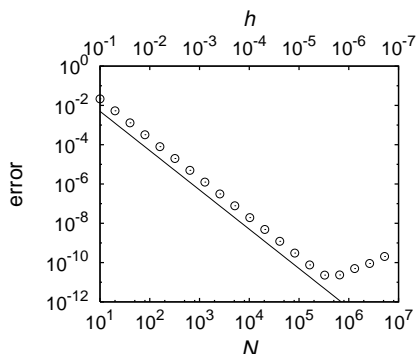


Figure 6-1: *Discretization and roundoff errors for a finite difference formula. The total error (circles), consisting of discretization and roundoff errors, decreases with resolution N until roundoff errors start to dominate. For comparison, the solid line shows the theoretical discretization error, proportional to $1/N^2$ or h^2 , with an arbitrary prefactor.*

To understand the plot more completely, we check for quantitative agreement. In the convergence test shown in the figure, double precision numbers are used with an accuracy of about 10^{-16} . The function values used for figure 1 are around 1, in order of magnitude, so that absolute and relative errors are approximately the same. The roundoff limitation occurs in this example at an accuracy of 10^{-11} . Why? In the formation of the difference $f(x+h) - f(x-h)$ a roundoff error of about 10^{-16} is introduced, but to obtain u , it is necessary to divide by $2h$, enhancing the error because h is a small number. In the figure the maximum accuracy is indeed approximately $10^{-16}/(2 \times 5 \times 10^{-6}) = 10^{-11}$. The total error is the sum of discretization error and roundoff error $O(h^2) + O(\epsilon/h)$, where $\epsilon \approx 10^{-16}$. The total error is a minimum when $h = O(\epsilon^{1/3}) = O(5 \times 10^{-6})$. This agrees perfectly with what is seen in figure 1.



Brainteaser: The convergence test indicates that $\|u_{2N} - u_N\| \rightarrow 0$ as the resolution N goes to infinity (roundoff ignored). Does this mean $\lim_{N \rightarrow \infty} \|u_N - u\| \rightarrow 0$, where u is the exact, correct answer?

6.3 Integration

The simplest way of numerical integration is to sum up function values. Rationale can be lent to this procedure by thinking of the function values

$f_j = f(x_j)$ as connected with straight lines. Let f_j denote the function at $x_j = x_0 + jh$. The area of the first trapezoidal segment is, using simple geometry, $(f_0 + f_1)h/2$. The area under the piecewise linear graph from x_0 to x_N is

$$\int_{x_0}^{x_N} f(x)dx \approx \frac{f_0 + f_1}{2}h + \frac{f_1 + f_2}{2}h + \dots = \left(\frac{f_0}{2} + f_1 + \dots + f_{N-1} + \frac{f_N}{2} \right) h,$$

which is indeed the sum of the function values. The boundary points carry only half the weight. This summation formula is called the “composite trapezoidal rule.”

Instead of straight lines it is also possible to *imagine* the function values are interpolated with quadratic polynomials. Fitting a parabola through three points and integrating, one obtains

$$\int_{x_0}^{x_2} f(x)dx \approx \frac{h}{3} (f_0 + 4f_1 + f_2).$$

For a parabola the approximate sign becomes an exact equality. This integration formula is well-known as “Simpson’s rule.” Repeated application of Simpson’s rule leads to

$$\int_{x_0}^{x_N} f(x)dx \approx \frac{h}{3} [f_0 + 4f_1 + 2f_2 + 4f_3 + 2f_4 + \dots + 4f_{N-1} + f_N].$$

An awkward feature about this “composite Simpson formula” is that function values are weighted unequally, although the grid points are equally spaced.

There is an exact relation between the integral and the sum of a function, known as “Euler-Maclaurin summation formula”:

$$\begin{aligned} \int_a^b f(x)dx &= h \sum_{j=1}^{N-1} f(a + jh) + \frac{h}{2} (f(a) + f(b)) + \\ &\quad - \sum_{j=1}^m h^{2j} \frac{B_{2j}}{(2j)!} (f^{(2j-1)}(b) - f^{(2j-1)}(a)) + \\ &\quad - h^{2m+2} \frac{B_{2m+2}}{(2m+2)!} (b-a) f^{(2m+2)}(\vartheta), \end{aligned}$$

where B_k are the Bernoulli numbers, $h = (b-a)/N$, and ϑ lies somewhere between a and b . The Bernoulli numbers are mathematical constants; the first few of them are $B_2 = 1/6$, $B_4 = -1/30$, $B_6 = 1/42$, \dots (Bernoulli numbers with odd indices are zero.) For $m = 0$,

$$\int_a^b f(x)dx = h \left(\frac{f_0}{2} + f_1 + \dots + f_{N-1} + \frac{f_N}{2} \right) - h^2 \frac{B_2}{2!} (b-a) f''(\vartheta).$$

The first order of the Euler-Maclaurin summation formula is the trapezoidal rule and the error for trapezoidal integration is $-h^2(b-a)f''(\vartheta)/12$.

The Euler-Maclaurin summation formula for $m = 1$ is

$$\begin{aligned} \int_a^b f(x)dx &= h \left(\frac{f_0}{2} + f_1 + \dots + f_{N-1} + \frac{f_N}{2} \right) + \\ &\quad -h^2 \frac{B_2}{2!} (f'(b) - f'(a)) - h^4 \frac{B_4}{4!} (b-a) f^{(4)}(\vartheta). \end{aligned}$$

There is no $O(h^3)$ error term. It is now apparent that the leading error in the composite trapezoidal rule arises from the boundaries only, not from the interior of the domain. If $f'(a)$ and $f'(b)$ are known or if they cancel each other, the integration error is only $-h^4(b-a)f^{(4)}(\vartheta)/720$.

The composite Simpson formula can be derived by using the Euler-Maclaurin summation formula with spacings h and $2h$. The integration error obtained in this way is $h^4(b-a) [\frac{1}{3}f^{(4)}(\vartheta_1) - \frac{4}{3}f^{(4)}(\vartheta_2)] /180$. Since the error is proportional to $f^{(4)}$, applying Simpson's rule to a *cubic* polynomial yields the integral exactly, although it is derived by integrating a *quadratic* polynomial.

The fourth-order error bound in the Simpson formula is larger than in the trapezoidal formula. The Simpson formula is only more accurate than the trapezoidal rule, because it better approximates the boundary regions. Away from the boundaries, the Simpson formula, the method of higher order, yields less accurate results than the trapezoidal rule, which is the penalty for the unequal coefficients. At the end, simple summation of function values is an excellent way of integration in the interior of the domain.



Recommended Reading: Abramovitz & Stegun, *Handbook of Mathematical Functions* includes a chapter on numerical analysis with finite difference, integration, and interpolation formulas and other helpful material. The book is also available online, at <http://www.math.sfu.ca/~cbm/aands/>