PHYS-4007/5007: Computational Physics
Course Lecture Notes
Section VI

Dr. Donald G. Luttermoser
East Tennessee State University

Version 7.0

Abstract

These class notes are designed for use of the instructor and students of the course PHYS-4007/5007: Computational Physics taught by Dr. Donald Luttermoser at East Tennessee State University.

VI. Numerical Differentiation and Integration

A. Derivatives.

1. The derivative of a function is defined by

$$\frac{df(x)}{dx} \equiv \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} \ . \tag{VI-1}$$

a) In computational work, functions and their independent variables are given by tabulated data and/or derived data.

b) Since there is a subtraction in Eq. (VI-1), subtraction cancellation can lead to rather large errors in the determination of a derivative via numerical techniques.

c) The computer's finite word length can cause the numerator to fluctuate between 0 and the machine precision ϵ_m as the denominator approaches zero.

d) In this section, we will discuss the various techniques used to calculate derivatives numerically and estimating the error in the derivative.

2. Method 1: Forward Difference.

a) Write out the function as a Taylor series at a position of *one step* forward from the current position:

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2!}f''(x) + \frac{h^3}{3!}f^{(3)}(x) + \cdots \ , \tag{VI-2}$$

where h is the step size as shown in Figure (VI-1).

b) We obtain the *forward-difference* derivative algorithm by solving Eq. (VI-1) for $f'(x)$:

$$f'_c(x) \ \simeq \ \frac{f(x+h) - f(x)}{h} \ , \tag{VI-3}$$

$$\simeq \; f'(x) + \frac{h}{2} f''(x) + \cdots \; , \qquad \text{(VI-4)}$$

where the subscript c denotes the computed expression.

i) The approximation of Eq. (VI-3) has an error proportional to h as shown in Eq. (VI-4).

ii) We can make the approximation error smaller and smaller by making h smaller and smaller.

iii) We can't make it too small however, since all precision will be lost if the subtraction cancellation error becomes larger than the step size.

c) Consider for example

$$f(x) = a + bx^2 \; . \qquad \text{(VI-5)}$$

the exact derivative is

$$f' = 2bx \; , \qquad \text{(VI-6)}$$

and the computed derivative is

$$f'_c(x) \approx \frac{f(x+h) - f(x)}{h} = 2bx + bh \; . \qquad \text{(VI-7)}$$

This would only be a good approximation if $h \ll 2x$.

3. Method 2: Central Difference.

a) An improved approximation to the derivative starts with the basic definition Eq. (VI-1). For this technique, instead of making a step of h forward, we form a *central difference* by stepping forward by $h/2$ and stepping backward by $h/2$:

$$\boxed{f'_c(x) \approx \frac{f(x+h/2) - f(x-h/2)}{h} \equiv D_c f(x,h) \; .} \qquad \text{(VI-8)}$$

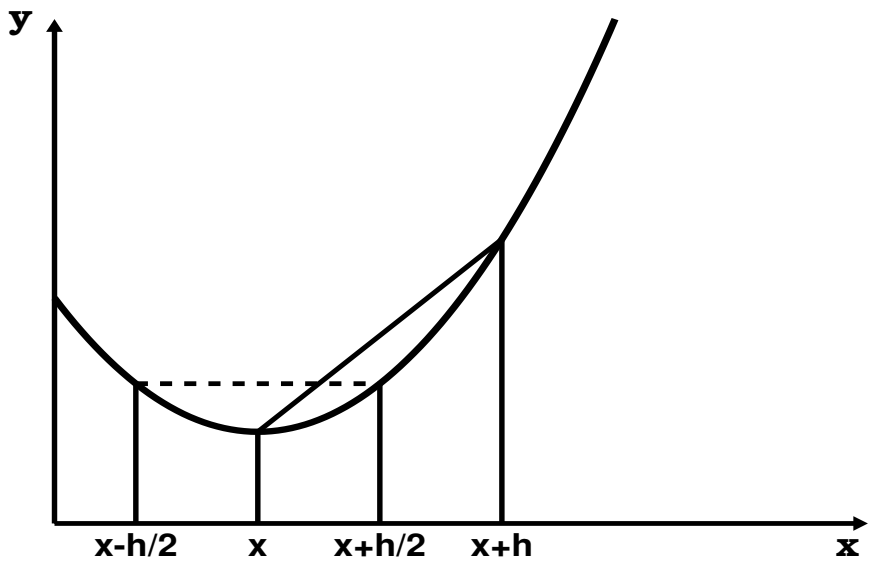

Figure VI–1: Forward difference (solid line) and central difference (dashed line) methods for numerical integration.

i) The symbol D_c represents center difference.

ii) Carrying out the Taylor series for $f(x \pm h)$ gives

$$f'_c \simeq f'(x) + \frac{1}{24} h^2 f^{(3)}(x) + \cdots \ . \qquad \text{(VI-9)}$$

iii) The important difference from Eq. (VI-3) is that when $f(x - h/2)$ is subtracted from $f(x + h/2)$, all terms containing an odd power of h in the Taylor series cancel.

iv) Therefore, the central-difference algorithm becomes accurate to one order higher than $h \Longrightarrow h^2$.

v) If $(f^{(3)} h^2)/24 \ll (f^{(2)} h)/2$, then the error with the central-difference method should be smaller than the forward difference.

b) For the polynomial given in Eq. (VI-5), the central difference gives the exact answer regardless of the size of h:

$$f_c'(x) \approx \frac{f(x+h/2) - f(x-h/2)}{h} = 2bx \ . \qquad \text{(VI-10)}$$

4. Errors in Taking Derivatives.

a) One should always try to keep calculation errors, ϵ_{tot}, at a minimum. This typically occurs when $\epsilon_{ro} \approx \epsilon_{\text{apprx}}$ in Eq. (IV-33).

b) Because differentiation subtracts 2 numbers that are close in value, the limit of roundoff error is essentially machine precision:

$$f'(x) \approx \frac{f(x+h) - f(x)}{h} \approx \frac{\epsilon_m}{h} \ , \qquad \text{(VI-11)}$$

$$\Longrightarrow \epsilon_{ro} \approx \frac{\epsilon_m}{h} \ . \qquad \text{(VI-12)}$$

c) The approximation error with the forward-difference algorithm (Eq. VI-4) is an $\mathcal{O}(h)$ term, while that with the central-difference algorithm (Eq. VI-9) is an $\mathcal{O}(h^2)$ term:

$$\epsilon_{\text{apprx}}^{\text{fd}} \approx \frac{f^{(2)}h}{2} \ , \qquad \text{(VI-13)}$$

$$\epsilon_{\text{apprx}}^{\text{cd}} \approx \frac{f^{(3)}h^2}{24} \ . \qquad \text{(VI-14)}$$

d) The value of h for which roundoff and approximation errors are equal is therefore

$$\frac{\epsilon_m}{h} \approx \epsilon_{\text{apprx}}^{\text{fd}} = \frac{f^{(2)}h}{2} \ , \qquad \text{(VI-15)}$$

$$\frac{\epsilon_m}{h} \approx \epsilon_{\text{apprx}}^{\text{cd}} = \frac{f^{(3)}h^2}{24} \ , \qquad \text{(VI-16)}$$

$$\Longrightarrow h_{\text{fd}}^2 = \frac{2\epsilon_m}{f^{(2)}} \ , \quad h_{\text{cd}}^3 = \frac{24\epsilon_m}{f^{(3)}} \ . \qquad \text{(VI-17)}$$

e) As an example, for the e^x and $\cos x$ functions $f' \approx f^{(2)} \approx f^{(3)}$ in single precision with $\epsilon_m \approx 10^{-7}$, one would get $h_{\rm fd} \approx 0.0005$ and $h_{\rm cd} \approx 0.01$. This makes the central-difference algorithm better for calculating derivatives since a larger h would mean a smaller error $\Longrightarrow$ here the error in the central-difference method is 20 times smaller than the error in the forward-difference method.

5. The Method in Calculating Second Derivatives.

a) Taking second derivatives will involve an additional subtraction step as compared to the first derivative calculation leading to a larger subtraction error.

b) We can remedy this with a little algebra in the central-difference method. Taking the second derivative of Eq. (VI-8) and then rewriting the first derivatives with a forward and backward difference equation, we get

$$f^{(2)} \simeq \frac{f'(x+h/2) - f'(x-h/2)}{h} \ , \qquad \text{(VI-18)}$$

$$\simeq \frac{1}{h^2} \left\{ [f(x+h/2) - f(x)] - [f(x) - f(x-h/2)] \right\} \ . \qquad \text{(VI-19)}$$

Note, however, that one must keep the second derivative formula in this form to minimize cancellation error.

B. Integration.

1. The method of numerical integration is sometimes referred to as **numerical quadrature** $\Longrightarrow$ summing boxes.

a) The definition of an integral given by Riemann is

$$\int_a^b f(x)\, dx = \lim_{h \to 0} \left[\sum_{i=1}^{(b-a)/h} f(x_i) \right] \ . \qquad \text{(VI-20)}$$

b) If we ignore the limit, the integral just becomes a summation of boxes or *quadrilaterals* lying below the curve:

$$\boxed{\int_a^b f(x)\,dx \approx \sum_{i=1}^{N} f(x_i)\,w_i\ ,} \qquad \text{(VI-21)}$$

where N is the number of points in the interval $[a, b]$ and f is evaluated at those interval points 'i', $f_i \equiv f(x_i)$. The w_i's are integration weights which are proportional to h the distance between points i and $i+1$.

c) The different integration schemes presented here will all make use of Eq. (VI-21).

d) In the simplest integration schemes, the integrand is approximated by a few terms in the Taylor series expansion of f and the terms are integrated $\Longrightarrow$ typically, adding more terms in the series yield higher precision.

e) This is the basis of the **Newton-Cotes** methods $\Longrightarrow$ the total interval is divided into equal subintervals with the integrand evaluated at equally spaced points x_i. Two such Newton-Cotes methods include the:

i) Trapezoid rule (a first-order method).

ii) Simpson rule (a second-order method).

f) More accurate integration methods involve the use of *non-equally spaced* intervals $\Longrightarrow$ these are the methods of **Gaussian quadrature**.

i) Gaussian quadrature methods are typically more precise than Newton-Cotes methods as long as there are no singularities (*i.e.*, denominators going to infinity, non-continuous functions) in the integrand or its derivative.

ii) You are better off to remove such singularities algebraically before attempting Gaussian quadrature. For example,

$$\int_{-1}^{1} |x|\, f(x)\, dx = \int_{-1}^{0} f(-x)\, dx + \int_{0}^{1} f(x)\, dx \ . \tag{VI-22}$$

iii) Regions where a function is slowly varying require fewer integration points, and regions with rapid variations in the function will have many integration points in order not to miss any oscillation — as can be seen for such functions, evenly spaced integration points will not represent the true integral accurately.

2. Trapezoid Rule.

a) As an introduction to the *numerically integration* of a function, consider the generic integral

$$I = \int_{a}^{b} f(x)\, dx \ . \tag{VI-23}$$

b) The most straight forward way to solving such an *integral function* is to evaluate $f(x)$ at a few points and fit a simple curve (*e.g.*, *piecewise linear*) through these points.

c) One way to do this is to fit trapezoids of equal width (h) under the curve represented by $f(x)$ and add up the total areas of these trapezoids. Let

$$h = \frac{b-a}{N-1} \ , \tag{VI-24}$$

$$x_i = a + (i-1)h \ , \quad i = 1, 2, ..., N \ . \tag{VI-25}$$

where a and b correspond to the initial and final endpoints and N the total number of points in the interval

$[a, b]$ (square bracket means we include the endpoints, *i.e.*, there are N–2 points in between the endpoints). Note that the trapezoid rule requires an *odd* number of points N.

i) Straight lines connect the points and this piecewise linear function serves as our fitting curve.

ii) The integral of this fitting function is easy to compute since it is the sum of the areas of trapezoids. The area of a single trapezoid is

$$T_i = \frac{1}{2}(x_{i+1} - x_i)(f_{i+1} + f_i) \ . \qquad \text{(VI-26)}$$

iii) The true integral is estimated as the sum of the areas of the trapezoids, so

$$I \approx I_T = T_1 + T_2 + \cdots + T_{N-1} = \sum_{i=1}^{N-1} T_i \ . \quad \text{(VI-27)}$$

Notice that the last term in the sum is $N - 1$ since there is one fewer panel than grid points.

iv) The general formula simplifies if we take equally spaced grid points as given by Eq. (VI-25).

v) Then the area for an individual trapezoid (*i.e.*, the area of the one trapezoid within that interval) becomes

$$T_i = \frac{1}{2}h(f_{i+1} + f_i) \ , \qquad \text{(VI-28)}$$

or in our original integral notation,

$$\int_{x_i}^{x_i+h} f(x)\, dx \simeq \frac{h(f_i + f_{i+1})}{2} = \frac{1}{2}hf_i + \frac{1}{2}hf_{i+1} \ . \qquad \text{(VI-29)}$$

As such, our weights in Eq. (VI-21) for the individual points are $w_i = \frac{1}{2}h$.

d) We now need to add up all of the trapezoids in the subintervals across the entire interval $[a, b] \Longrightarrow$ the **trapezoid rule**:

$$\begin{aligned} I_T(h) &= \frac{1}{2}hf_1 + hf_2 + hf_3 + \cdots + hf_{N-1} + \frac{1}{2}hf_N \\ &= \frac{1}{2}h(f_1 + f_N) + h\sum_{i=2}^{N-1} f_i \end{aligned} \qquad \text{(VI-30)}$$

or

$$\boxed{\int_a^b f(x)\,dx \approx \frac{h}{2}f_1 + hf_2 + hf_3 + \cdots + hf_{N-1} + \frac{h}{2}f_N \ .} \qquad \text{(VI-31)}$$

e) Note that since each internal point gets counted twice, it has a weight of h, whereas the endpoints get counted just once and thus have weights of $h/2$:

$$w_i = \left\{\frac{h}{2}, h, ..., h, \frac{h}{2}\right\} \ . \qquad \text{(VI-32)}$$

f) Our **approximation error**, also called the *truncation error* or the *algorithmic error* here, for the trapezoid rule can be written either as

$$\begin{aligned} \epsilon_{\rm apprx} &= I - I_T(x, h) \\ &= -\frac{1}{12}(b-a)h^2 f''(\zeta) \end{aligned} \qquad \text{(VI-33)}$$

for some value $x = \zeta$ that lies in $[a, b]$ or as

$$\epsilon_{\rm apprx} = -\frac{1}{12}h^2[f'(b) - f'(a)] + \mathcal{O}(h^4) \ . \qquad \text{(VI-34)}$$

g) This error is proportional to h^2 and Eq. (VI-34) warns you that the trapezoidal rule will have difficulties if the derivative diverges at the end points.

3. Simpson's Rule ("Simpson, eh?" – M. Burns).

a) Instead of fitting two adjacent points with trapezoids, we will now fit three adjacent points with parabolas:

$$f(x) \approx \alpha x^2 + \beta x + \gamma \ , \tag{VI-35}$$

for each interval, still keeping the intervals evenly spaced.

b) The area of each section is then the integral of this parabola:

$$\int_{x_i}^{x_i+h} (\alpha x^2 + \beta x + \gamma)\, dx = \frac{\alpha x^3}{3} + \frac{\beta x^2}{2} + \gamma x \Big|_{x_i}^{x_i+h} \ . \tag{VI-36}$$

c) This is equivalent to integrating the Taylor series up to the quadratic term. Hence the Simpson rule is a second-order polynomial method.

d) In order to relate the parameters α, β, and γ to the function, we consider an interval from -1 to $+1$,

$$\int_{-1}^{1} (\alpha x^2 + \beta x + \gamma)\, dx = \frac{2\alpha}{3} + 2\gamma \ . \tag{VI-37}$$

i) Note, however, for the function itself,

$$\begin{gathered} f(-1) = \alpha - \beta + \gamma \ , \quad \alpha = \tfrac{f(1)+f(-1)}{2} - f(0) \ , \\ f(0) = \gamma \ , \quad \beta = \tfrac{f(1)-f(-1)}{2} \ , \\ f(1) = \alpha + \beta + \gamma \ , \quad \gamma = f(0) \ . \end{gathered} \tag{VI-38}$$

ii) Using the results of Eq. (VI-38) in Eq. (VI-37) yields,

$$\int_{-1}^{1} (\alpha x^2 + \beta x + \gamma)\, dx = \frac{f(-1)}{3} + \frac{4f(0)}{3} + \frac{f(1)}{3} \ . \tag{VI-39}$$

iii) Because 3 values of the function are needed, we evaluate the integral over two adjacent intervals $\Longrightarrow$ evaluate the functions at the two endpoints and the middle:

$$\begin{aligned} \int_{x_i-h}^{x_i+h} f(x)\,dx &= \int_{x_i}^{x_i+h} f(x)\,dx + \int_{x_i-h}^{x_i} f(x)\,dx \\ &\simeq \frac{h}{3} f_{i-1} + \frac{4h}{3} f_i + \frac{h}{3} f_{i+1} \ . \end{aligned} \tag{VI-40}$$

e) Simpson's rule requires the elementary integration to be over *pairs* of intervals $\Longrightarrow$ this requires the number of intervals to be even, and hence, the number of points N to be odd.

f) To integrate across the entire range $[a, b]$, we add up contributions from each pair of subintervals, counting all but the first and last endpoints twice:

$$\boxed{\begin{aligned} \textstyle\int_a^b f(x)\,dx \approx\ & \tfrac{h}{3} f_1 + \tfrac{4h}{3} f_2 + \tfrac{2h}{3} f_3 + \tfrac{4h}{3} f_4 + \cdots \\ & + \tfrac{4h}{3} f_{N-1} + \tfrac{h}{3} f_N \ . \end{aligned}} \tag{VI-41}$$

i) From this integral, we see that the total set of weights is

$$w_i = \left\{ \frac{h}{3}, \frac{4h}{3}, \frac{2h}{3}, \frac{4h}{3}, ..., \frac{2h}{3}, \frac{4h}{3}, \frac{h}{3} \right\} \ . \tag{VI-42}$$

ii) The sum of your weights provides a useful check on your integration:

$$\sum_{i=1}^{N} w_i = (N-1)\,h \ . \tag{VI-43}$$

Remember, N must be odd.

4. Errors in the Newton-Cotes Methods.

a) As we have said above, and as was the case for differentiation, our Newton-Cotes methods are essentially just Taylor series expansions of a function. We will highlight here what is derived in more detail in the textbook.

b) The *approximation* (*i.e.*, *truncation* or *algorithmic*) error ϵ_{apprx} can be estimated by the next higher term in the Taylor series that is not used in the evaluation of the integral:

$$\epsilon_{\text{apprx-t}} = \mathcal{O}\left(\frac{[b-a]^3}{N^2}\right) f^{(2)} , \qquad \text{(VI-44)}$$

$$\epsilon_{\text{apprx-s}} = \mathcal{O}\left(\frac{[b-a]^5}{N^4}\right) f^{(4)} , \qquad \text{(VI-45)}$$

where the subscripts t and s in the subscripts refer to *trapezoid* and *Simpson's* rule, respectively.

c) The *relative error* ϵ is just these approximation errors divided by the value of the function:

$$\epsilon_{t,s} = \frac{\epsilon_{\text{apprx-t,s}}}{f} . \qquad \text{(VI-46)}$$

d) Assume that after N steps, the relative roundoff error is random, so

$$\epsilon_{ro} \approx \sqrt{N}\, \epsilon_m , \qquad \text{(VI-47)}$$

where ϵ_m is the machine precision ($\sim 10^{-7}$ for single precision, 10^{-15} for double precision).

e) Assume that the total error is given by Eq. (IV-33):

$$\epsilon_{\text{tot}} = \epsilon_{ro} + \epsilon_{\text{apprx}}. \qquad \text{(VI-48)}$$

i) We want to determine the value of N that minimizes the total error. This will occur when the

two relative errors are of equal magnitude, which we approximate even further by assuming that the two errors are equal:

$$\epsilon_{ro} = \epsilon_{\rm t,s} = \frac{\epsilon_{\rm apprx\text{-}t,s}}{f} \ . \qquad \text{(VI-49)}$$

ii) Now, let's make the following assumptions:

$$\frac{f^{(n)}}{f} \approx 1 \ , \qquad \text{(VI-50)}$$

$$b - a = 1 \implies h = \frac{1}{N} \ . \qquad \text{(VI-51)}$$

f) For the **trapezoid rule**, Eq. (VI-49) becomes

$$\sqrt{N}\,\epsilon_m \approx \frac{f^{(2)}(b-a)^3}{fN^2} = \frac{1}{N^2} \ , \qquad \text{(VI-52)}$$

$$\Rightarrow N \approx \frac{1}{(\epsilon_m)^{2/5}} \ . \qquad \text{(VI-53)}$$

i) Then, the optimum numbers of N steps for the trapezoid rule are

$$N = \frac{1}{h} = \begin{cases} (1/10^{-7})^{2/5} = 631 \ , & \text{for single precision,} \\ (1/10^{-15})^{2/5} = 10^6 \ , & \text{for double precision.} \end{cases} \qquad \text{(VI-54)}$$

ii) The corresponding errors are

$$\epsilon_{ro} \approx \sqrt{N}\,\epsilon_m = \begin{cases} 3 \times 10^{-6} \ , & \text{for single precision,} \\ 10^{-12} \ , & \text{for double precision.} \end{cases} \qquad \text{(VI-55)}$$

g) For **Simpson's rule**, Eq. (VI-49) becomes

$$\sqrt{N}\,\epsilon_m \approx \frac{f^{(4)}(b-a)^5}{fN^4} = \frac{1}{N^4} \ , \qquad \text{(VI-56)}$$

$$\Rightarrow N \approx \frac{1}{(\epsilon_m)^{2/9}} \ . \qquad \text{(VI-57)}$$

i) Then, the optimum numbers of N steps for Simpson's rule are

$$N = \frac{1}{h} = \begin{cases} (1/10^{-7})^{2/9} = 36 \ , & \text{for single precision,} \\ (1/10^{-15})^{2/9} = 2154 \ , & \text{for double precision.} \end{cases} \tag{VI-58}$$

ii) The corresponding errors are

$$\epsilon_{ro} \approx \sqrt{N}\,\epsilon_m = \begin{cases} 6 \times 10^{-7} \ , & \text{for single precision,} \\ 5 \times 10^{-14} \ , & \text{for double precision.} \end{cases} \tag{VI-59}$$

h) These results are illuminating because they show that

i) Simpson's rule is an improvement over the trapezoid rule.

ii) It is possible to obtain an error close to the machine precision with Simpson's rule (and with other higher-order integration algorithms).

iii) Obtaining the best numerical approximation to an integral is not obtained by letting $N \to \infty$, but with a relatively small $N \leq 1000$.

5. There are higher order Newton-Cotes methods, the 3rd-degree "3/8th" method and the 4th-degree Milne method. See Table 5.1 in your textbook on page 61 for the elementary weights that one would use for these two methods.

6. See Appendix G for detailed information of the **Gaussian Quadrature** technique of numerical integration.