

The vector  $\mathbf{x}_0$  accepts the initial guesses for  $A$  and  $B$  and is updated as  $A$  and  $B$  are modified through an iteration procedure to converge to the least-square values. Note that the sum is simply a statement of the quantity to be minimized, namely (3.3.5). The results of this calculation can be illustrated with MATLAB *figure 7*

```
xga=-3:0.1:3;
a=coeff(1); b=coeff(2)
yga=a*exp(-b*xga.^2);
figure(7), plot(x2,y2,'0',xga,yga,'m')
```

Note that for this case, the initial guess is extremely important. For any given problem where this technique is used, an educated guess for the values of parameters like  $A$  and  $B$  can determine if the technique will work at all. The results should also be checked carefully since there is no guarantee that a minimization near the desired fit can be found.

## 4 Numerical Differentiation and Integration

Differentiation and integration form the backbone of the mathematical techniques required to describe and analyze physical systems. These two mathematical concepts describe how certain quantities of interest change with respect to either space and time or both. Understanding how to evaluate these quantities numerically is essential to understanding systems beyond the scope of analytic methods.

### 4.1 Numerical Differentiation

Given a set of data or a function, it may be useful to differentiate the quantity considered in order to determine a physically relevant property. For instance, given a set of data which represents the position of a particle as a function of time, then the derivative and second derivative give the velocity and acceleration respectively. From calculus, the definition of the derivative is given by

$$\frac{df(t)}{dt} = \lim_{\Delta t \rightarrow 0} \frac{f(t + \Delta t) - f(t)}{\Delta t} \quad (4.1.1)$$

Since the derivative is the slope, the formula on the right is nothing more than a rise-over-run formula for the slope. The general idea of calculus is that as  $\Delta t \rightarrow 0$ , then the rise-over-run gives the instantaneous slope. Numerically, this means that if we take  $\Delta t$  sufficiently small, then the approximation should be fairly accurate. To quantify and control the error associated with approximating the derivative, we make use of *Taylor series* expansions.

To see how the Taylor expansions are useful, consider the following two Taylor series:

$$f(t + \Delta t) = f(t) + \Delta t \frac{df(t)}{dt} + \frac{\Delta t^2}{2!} \frac{d^2 f(t)}{dt^2} + \frac{\Delta t^3}{3!} \frac{d^3 f(c_1)}{dt^3} \quad (4.1.2a)$$

$$f(t - \Delta t) = f(t) - \Delta t \frac{df(t)}{dt} + \frac{\Delta t^2}{2!} \frac{d^2 f(t)}{dt^2} - \frac{\Delta t^3}{3!} \frac{d^3 f(c_2)}{dt^3} \quad (4.1.2b)$$

where  $c_i \in [a, b]$ . Subtracting these two expressions gives

$$f(t + \Delta t) - f(t - \Delta t) = 2\Delta t \frac{df(t)}{dt} + \frac{\Delta t^3}{3!} \left( \frac{d^3 f(c_1)}{dt^3} + \frac{d^3 f(c_2)}{dt^3} \right). \quad (4.1.3)$$

By using the mean-value theorem of calculus, we find  $f'''(c) = (f'''(c_1) + f'''(c_2))/2$ . Upon dividing the above expression by  $2\Delta t$  and rearranging, we find the following expression for the first derivative:

$$\frac{df(t)}{dt} = \frac{f(t + \Delta t) - f(t - \Delta t)}{2\Delta t} - \frac{\Delta t^2}{6} \frac{d^3 f(c)}{dt^3} \quad (4.1.4)$$

where the last term is the truncation error associated with the approximation of the first derivative using this particular Taylor series generated expression. Note that the truncation error in this case is  $O(\Delta t^2)$ .

We could improve on this by continuing our Taylor expansion and truncating it at higher orders in  $\Delta t$ . This would lead to higher accuracy schemes. Specifically, by truncating at  $O(\Delta t^5)$ , we would have

$$f(t + \Delta t) = f(t) + \Delta t \frac{df(t)}{dt} + \frac{\Delta t^2}{2!} \frac{d^2 f(t)}{dt^2} + \frac{\Delta t^3}{3!} \frac{d^3 f(t)}{dt^3} + \frac{\Delta t^4}{4!} \frac{d^4 f(t)}{dt^4} + \frac{\Delta t^5}{5!} \frac{d^5 f(c_1)}{dt^5} \quad (4.1.5a)$$

$$f(t - \Delta t) = f(t) - \Delta t \frac{df(t)}{dt} + \frac{\Delta t^2}{2!} \frac{d^2 f(t)}{dt^2} - \frac{\Delta t^3}{3!} \frac{d^3 f(t)}{dt^3} + \frac{\Delta t^4}{4!} \frac{d^4 f(t)}{dt^4} - \frac{\Delta t^5}{5!} \frac{d^5 f(c_2)}{dt^5} \quad (4.1.5b)$$

where  $c_i \in [a, b]$ . Again subtracting these two expressions gives

$$f(t + \Delta t) - f(t - \Delta t) = 2\Delta t \frac{df(t)}{dt} + \frac{2\Delta t^3}{3!} \frac{d^3 f(t)}{dt^3} + \frac{\Delta t^5}{5!} \left( \frac{d^5 f(c_1)}{dt^5} + \frac{d^5 f(c_2)}{dt^5} \right). \quad (4.1.6)$$

In this approximation, there is third derivative term left over which needs to be removed. By using two additional points to approximate the derivative, this term can be removed. Thus we use the two additional points  $f(t + 2\Delta t)$  and  $f(t - 2\Delta t)$ . Upon replacing  $\Delta t$  by  $2\Delta t$  in (4.1.6), we find

$$f(t + 2\Delta t) - f(t - 2\Delta t) = 4\Delta t \frac{df(t)}{dt} + \frac{16\Delta t^3}{3!} \frac{d^3 f(t)}{dt^3} + \frac{32\Delta t^5}{5!} \left( \frac{d^5 f(c_3)}{dt^5} + \frac{d^5 f(c_4)}{dt^5} \right). \quad (4.1.7)$$

---



---

$O(\Delta t^2)$  center-difference schemes

---



---

$$\begin{aligned} f'(t) &= [f(t + \Delta t) - f(t - \Delta t)]/2\Delta t \\ f''(t) &= [f(t + \Delta t) - 2f(t) + f(t - \Delta t)]/\Delta t^2 \\ f'''(t) &= [f(t + 2\Delta t) - 2f(t + \Delta t) + 2f(t - \Delta t) - f(t - 2\Delta t)]/2\Delta t^3 \\ f''''(t) &= [f(t + 2\Delta t) - 4f(t + \Delta t) + 6f(t) - 4f(t - \Delta t) + f(t - 2\Delta t)]/\Delta t^4 \end{aligned}$$


---



---

Table 4: Second-order accurate center-difference formulas.

By multiplying (4.1.6) by eight and subtracting (4.1.7) and using the mean-value theorem on the truncation terms twice, we find the expression:

$$\frac{df(t)}{dt} = \frac{-f(t + 2\Delta t) + 8f(t + \Delta t) - 8f(t - \Delta t) + f(t - 2\Delta t)}{12\Delta t} + \frac{\Delta t^4}{30} f^{(5)}(c) \quad (4.1.8)$$

where  $f^{(5)}$  is the fifth derivative and the truncation is of  $O(\Delta t^4)$ .

Approximating higher derivatives works in a similar fashion. By starting with the pair of equations (4.1.2) and adding, this gives the result

$$f(t + \Delta t) + f(t - \Delta t) = 2f(t) + \Delta t^2 \frac{d^2 f(t)}{dt^2} + \frac{\Delta t^4}{4!} \left( \frac{d^4 f(c_1)}{dt^4} + \frac{d^4 f(c_2)}{dt^4} \right). \quad (4.1.9)$$

By rearranging and solving for the second derivative, the  $O(\Delta t^2)$  accurate expression is derived

$$\frac{d^2 f(t)}{dt^2} = \frac{f(t + \Delta t) - 2f(t) + f(t - \Delta t)}{\Delta t^2} + O(\Delta t^2) \quad (4.1.10)$$

where the truncation error is of  $O(\Delta t^2)$  and is found again by the mean-value theorem to be  $-(\Delta t^2/12)f''''(c)$ . This process can be continued to find any arbitrary derivative. Thus, we could also approximate the third, fourth, and higher derivatives using this technique. It is also possible to generate backward and forward difference schemes by using points only behind or in front of the current point respectively. Tables 8-10 summarize the second-order and fourth-order central difference schemes along with the forward- and backward-difference formulas which are accurate to second-order.

A final remark is in order concerning these differentiation schemes. The central difference schemes are an excellent method for generating the values of the derivative in the interior points of a data set. However, at the end points, forward and backward difference methods must be used since they do not have

---



---

$O(\Delta t^4)$  center-difference schemes

---



---

$$\begin{aligned}
 f'(t) &= [-f(t+2\Delta t) + 8f(t+\Delta t) - 8f(t-\Delta t) + f(t-2\Delta t)]/12\Delta t \\
 f''(t) &= [-f(t+2\Delta t) + 16f(t+\Delta t) - 30f(t) \\
 &\quad + 16f(t-\Delta t) - f(t-2\Delta t)]/12\Delta t^2 \\
 f'''(t) &= [-f(t+3\Delta t) + 8f(t+2\Delta t) - 13f(t+\Delta t) \\
 &\quad + 13f(t-\Delta t) - 8f(t-2\Delta t) + f(t-3\Delta t)]/8\Delta t^3 \\
 f''''(t) &= [-f(t+3\Delta t) + 12f(t+2\Delta t) - 39f(t+\Delta t) + 56f(t) \\
 &\quad - 39f(t-\Delta t) + 12f(t-2\Delta t) - f(t-3\Delta t)]/6\Delta t^4
 \end{aligned}$$


---



---

Table 5: Fourth-order accurate center-difference formulas.

---



---

$O(\Delta t^2)$  forward- and backward-difference schemes

---



---

$$\begin{aligned}
 f'(t) &= [-3f(t) + 4f(t+\Delta t) - f(t+2\Delta t)]/2\Delta t \\
 f'(t) &= [3f(t) - 4f(t-\Delta t) + f(t-2\Delta t)]/2\Delta t \\
 f''(t) &= [2f(t) - 5f(t+\Delta t) + 4f(t+2\Delta t) - f(t+3\Delta t)]/\Delta t^3 \\
 f''(t) &= [2f(t) - 5f(t-\Delta t) + 4f(t-2\Delta t) - f(t-3\Delta t)]/\Delta t^3
 \end{aligned}$$


---



---

Table 6: Second-order accurate forward- and backward-difference formulas.

neighboring points to the left and right respectively. Thus special care must be taken at the end points of any computational domain.

It may be tempting to deduce from the difference formulas that as  $\Delta t \rightarrow 0$ , the accuracy only improves in these computational methods. However, this line of reasoning completely neglects the second source of error in evaluating derivatives: numerical round-off.

### Round-off and optimal step-size

An unavoidable consequence of working with numerical computations is round-off error. When working with most computations, *double precision* numbers are used. This allows for 16-digit accuracy in the representation of a given number. This round-off has significant impact upon numerical computations

and the issue of time-stepping.

As an example of the impact of round-off, we consider the approximation to the derivative

$$\frac{dy}{dt} \approx \frac{y(t + \Delta t) - y(t)}{\Delta t} + \epsilon(y(t), \Delta t) \quad (4.1.11)$$

where  $\epsilon(y(t), \Delta t)$  measures the truncation error. Upon evaluating this expression in the computer, round-off error occurs so that

$$y(t) = Y(t) + e(t), \quad (4.1.12)$$

where  $Y(t)$  is the approximated value given by the computer and  $e(t)$  measures the error from the true value  $y(t)$ . Thus the combined error between the round-off and truncation gives the following expression for the derivative:

$$\frac{dy}{dt} = \frac{y(t + \Delta t) - y(t)}{\Delta t} + E(y(t), \Delta t) \quad (4.1.13)$$

where the total error,  $E$ , is the combination of round-off and truncation such that

$$E = E_{\text{round}} + E_{\text{trunc}} = \frac{e(t + \Delta t) - e(t)}{\Delta t} - \frac{\Delta t^2}{2} \frac{d^2 y(c)}{dt^2}. \quad (4.1.14)$$

We now determine the maximum size of the error. In particular, we can bound the maximum value of round-off and the derivate to be

$$|e(t + \Delta t)| \leq e_r \quad (4.1.15a)$$

$$|-e(t)| \leq e_r \quad (4.1.15b)$$

$$M = \max_{c \in [t_n, t_{n+1}]} \left\{ \left| \frac{d^2 y(c)}{dt^2} \right| \right\}. \quad (4.1.15c)$$

This then gives the maximum error to be

$$|E| \leq \frac{e_r + e_r}{\Delta t} + \frac{\Delta t^2}{2} M = \frac{2e_r}{\Delta t} + \frac{\Delta t^2 M}{2}. \quad (4.1.16)$$

Note that as  $\Delta t$  gets large, the error grows quadratically due to the truncation error. However, as  $\Delta t$  decreases to zero, the error is dominated by round-off which grows like  $1/\Delta t$ .

To minimize the error, we require that  $\partial|E|/\partial(\Delta t) = 0$ . Calculating this derivative gives

$$\frac{\partial|E|}{\partial(\Delta t)} = -\frac{2e_r}{\Delta t^2} + M\Delta t = 0, \quad (4.1.17)$$

so that

$$\Delta t = \left( \frac{2e_r}{M} \right)^{1/3}. \quad (4.1.18)$$

This gives the step size resulting in a minimum error. Thus the smallest step-size is not necessarily the most accurate. Rather, a balance between round-off error and truncation error is achieved to obtain the optimal step-size. For  $e_r \approx 10^{-16}$ , the optimal  $\Delta t \approx 10^{-5}$ . Below this value of  $\Delta t$ , numerical round-off begins to dominate the error

A similar procedure can be carried out for evaluating the optimal step size associated with the  $O(\Delta t^4)$  accurate scheme for the first derivative. In this case

$$\frac{dy}{dt} = \frac{-f(t+2\Delta t) + 8f(t+\Delta t) - 8f(t-\Delta t) + f(t-2\Delta t)}{12\Delta t} + E(y(t), \Delta t) \quad (4.1.19)$$

where the total error,  $E$ , is the combination of round-off and truncation such that

$$E = \frac{-e(t+2\Delta t) + 8e(t+\Delta t) - 8e(t-\Delta t) + e(t-2\Delta t)}{12\Delta t} + \frac{\Delta t^4}{30} \frac{d^5 y(c)}{dt^5}. \quad (4.1.20)$$

We now determine the maximum size of the error. In particular, we can bound the maximum value of round-off to  $e$  as before and set  $M = \max\{|y''''(c)|\}$ . This then gives the maximum error to be

$$|E| = \frac{3e_r}{2\Delta t} + \frac{\Delta t^4 M}{30}. \quad (4.1.21)$$

Note that as  $\Delta t$  gets large, the error grows like a quartic due to the truncation error. However, as  $\Delta t$  decreases to zero, the error is again dominated by round-off which grows like  $1/\Delta t$ .

To minimize the error, we require that  $\partial|E|/\partial(\Delta t) = 0$ . Calculating this derivative gives

$$\Delta t = \left(\frac{45e_r}{4M}\right)^{1/5}. \quad (4.1.22)$$

Thus in this case, the optimal step  $\Delta t \approx 10^{-3}$ . This shows that the error can be quickly dominated by numerical round-off if one is not careful to take this significant effect into account.