

Note on the Numerical Evaluation of a First Derivative from a Table of a Function Satisfying a Second Order Differential Equation

By J. C. P. Miller

It is well known that the evaluation of derivatives from a table of equally-spaced function values and differences is an unsatisfactory process. Differentiation is a "local" process, and any rounding errors produce uncertainties which are enhanced with small intervals in the tabulation. On the other hand, the initial convergence of the series of multiples of differences that represent a formula for a derivative is improved by using small intervals h . The best interval to use for numerical differentiation is usually not that given in the table, but one which produces a series which only just "converges," i.e. one where the last term available is just on the verge of being negligible.

Kopal (1955) discusses a simplified version of this problem, and shows how to develop a criterion for choice of interval to produce derivatives of accuracy as great as is convenient, though he restricts the choice of method considered, for reasons of clarity in exposition, and in order to obtain a definite conclusion.

In the present note we consider one particular case where better results are obtainable. This is when a function is tabulated at some interval h , the function being one that satisfies a second order differential equation

$$y'' = f(x, y)$$

in which we suppose, for the moment, that the first derivative is absent. There is then no difficulty about obtaining an accurate second derivative; we simply substitute the adequately known value of y in the differential equation. Now, integration is a stable process, in the sense that errors in an integral tend to be relatively reduced with respect to those in the integrand (whereas those in a derivative are relatively larger). We may, therefore, integrate y'' and obtain good values of y' , *except for a badly determined constant of integration*.

We may, however, integrate again, equally accurately, and recover y , but with two undetermined (or relatively so) constants of integration. The suggested method for obtaining the constants is to use two fairly widely-spaced values of y , *taken from the tables*, to give equations for these constants. The "y" constant can be determined to an accuracy comparable with that of the tabular values y , but is not needed; the other, "y," constant is determined with a precision that increases with the interval between the two chosen values of y .

The initial derivation of the formula, in spite of its evident possibilities of symmetry, was rather clumsy; we will, therefore, simply quote the formula with explanations of its various parts, and give a formal verification using operators.

We suppose $y_r = y(a + rh)$ given for $\pm r = 0, 1, 2, 3, \dots$ where $y(x)$ satisfies the differential equation

$y'' = f(x, y)$ and that we desire an accurate value of $y'(a)$. Such a value is given by the formula

$$\begin{aligned} hy'(a) = & \frac{1}{2n}(y_n - y_{-n}) - \frac{h^2}{2n} \sum_1^{n-1} (n-r)(f_r - f_{-r}) \\ & - \frac{h^2}{2n} \left\{ \frac{1}{12} - \frac{1}{240}\delta^2 + \frac{31}{60480}\delta^4 - \dots \right\} (f_n - f_{-n}) \\ & - h^2 \left\{ \frac{1}{12}\mu\delta - \frac{11}{720}\mu\delta^3 + \frac{191}{60480}\mu\delta^5 - \dots \right\} f_0. \end{aligned}$$

The first term of the right uses the difference between widely separated values of y_r , whence comes the accuracy in determining the "y'" constant of integration. The weighted sum represents a double sum of the second derivative which is converted into a double integral by the correction terms in the second line. The sum has a jump in the middle, which is needed to preserve symmetry and which, with the help of the corrections in the last line produces a first derivative of y . This identification of terms is somewhat crude, but may be helpful.

We verify by use of operators D , E and δ , detaching the operand f_0 in all cases.

The sum operator on the right is

$$\sum_1^{n-1} (n-r)(E^r - E^{-r}).$$

Now

$$\begin{aligned} \sum_1^{n-1} (n-r)E^r &= E^{n-1} + 2E^{n-2} + 3E^{n-3} + \dots + (n-1)E \\ &= \frac{E^{n-1}}{(1-E^{-1})^2} - \frac{n}{1-E^{-1}} - \frac{E^{-1}}{(1-E^{-1})^2} \\ &= \frac{E^n - 1}{\delta^2} - \frac{nE}{E-1} \end{aligned}$$

since $(E-1)(1-E^{-1}) = \delta^2$.

Thus the complete sum operator is

$$\begin{aligned} & \left(\frac{E^n - E^{-n}}{\delta^2} - \frac{nE}{E-1} + \frac{nE^{-1}}{E^{-1}-1} \right) D^2 \\ &= \left(\frac{E^n - E^{-n}}{\delta^2} - \frac{2n\mu}{\delta} \right) D^2. \end{aligned}$$

Again, we have the familiar expressions for single and double integrals

$$\begin{aligned} \frac{1}{hD} &= \frac{\mu}{\delta} \left(1 - \frac{1}{12}\delta^2 + \frac{11}{720}\delta^4 - \frac{191}{60480}\delta^6 + \dots \right) \\ \frac{1}{h^2 D^2} &= \frac{1}{\delta^2} \left(1 + \frac{1}{12}\delta^2 - \frac{1}{240}\delta^4 + \frac{31}{60480}\delta^6 - \dots \right). \end{aligned}$$

Evaluation of First Derivative

x	Bix	$\text{Bi}''x$	δ	δ^2	δ^3	δ^4	δ^5
-1.0	0.103997	-0.103997		+12859		-319	
-0.9	0.162639	0.146375					
-0.8	0.219828	0.175862					
-0.7	0.275268	0.192688					
-0.6	0.328792	0.197275					
-0.5	0.380353	-0.190176		+11069		-5	
-0.4	0.430021	0.172008					
-0.3	0.477978	0.143393					
-0.2	0.524509	0.104902					
-0.1	0.569999	0.057000					
0.0	0.614927	0.000000	+57000		-112		+46
+0.1	0.659862	+0.065986	+65986		+135		+43
+0.2	0.705464	0.141093					
+0.3	0.752486	0.225746					
+0.4	0.801773	0.320709					
+0.5	0.854277	+0.427138		+13071		+516	
+0.6	0.911063	0.546638					
+0.7	0.973329	0.681330					
+0.8	1.042422	0.833938					
+0.9	1.119873	1.007886					
+1.0	1.207424	+1.207424		+30819		+1175	

See, for instance, *Interpolation and Applied Tables* (1956), p. 68.

The complete operator on the right is then

$$\begin{aligned} & \frac{1}{2n}(E^n - E^{-n}) - \frac{h^2 D^2}{2n} \left(\frac{E^n - E^{-n}}{\delta^2} - \frac{2n\mu}{\delta} \right) \\ & - \frac{h^2 D^2}{2n} \left(\frac{1}{h^2 D^2} - \frac{1}{\delta^2} \right) (E - E^{-n}) - h^2 D^2 \left(\frac{\mu}{\delta} - \frac{1}{hD} \right) \\ & = (E^n - E^{-n}) \left(\frac{1}{2n} - \frac{h^2 D^2}{2n\delta^2} - \frac{1}{2n} + \frac{h^2 D^2}{2n^2} \right) \\ & + h^2 D^2 \left(\frac{\mu}{\delta} - \frac{\mu}{\delta} + \frac{1}{hD} \right) = hD. \end{aligned}$$

The formula is thus verified.

If the differential equation is one in which the first derivative is present, the position is more complicated.

If the equation is linear we may, by change of dependent variable alone, reduce the equation to normal form, with y' absent, and still retain the same independent variable x , and so also the original tabular interval. In this case, therefore, we can still use the formula.

Other cases are also possible, but the more obvious approaches involve a certain amount of trial and error, and the investigation will not be carried further in this note.

References

- KOPAL, Z. (1955). *Numerical Analysis*, Chapter III—§ E. Chapman and Hall.
 MILLER, J. C. P. (1946). *The Airy Integral*, B.A. Mathematical Tables, Part-Volume B. Cambridge, University Press.
 H.M. NAUTICAL ALMANAC OFFICE (1956). *Interpolation and Allied Tables*. London, H.M. Stationery Office.

Numerical Example

Consider the equation for the Airy Integral

$$y'' = xy$$

and use the table of Bix [see Miller (1946), page B44] for $x = -1(0.1) + 1$ to obtain two values of $\text{Bi}(0)$, with $n = 5$ and $n = 10$ respectively. Values to 6 decimals are listed, with differences used in the formula, in the table above. With $h = 0.1$, $n = 5$, the terms of the formula at the top of p. 112 give

$$\begin{aligned} \frac{1}{10} \text{Bi}'(0) &= 0.0473924,0 - 0.0024609,2 \\ &\quad - 0.000514,3 - 0.0000512,4 \end{aligned}$$

whence $\text{Bi}'(0) = 0.448288,1$.

With $n = 10$, the formula gives

$$\begin{aligned} \frac{1}{10} \text{Bi}'(0) &= 0.0551713,5 - 0.0102366,2 \\ &\quad - 0.0000546,1 - 0.0000512,4 \end{aligned}$$

whence $\text{Bi}'(0) = 0.448288,8$. The true value is 0.44828836.

Almost all the error comes from the first term and may amount to $1/n$ times the maximum error in Bix , that is to 10 units in the last place given (the commas are used to indicate that this is a guard figure) when $n = 5$ or 5 units when $n = 10$. The other terms contribute hardly more than rounding errors in the guard figure; in fact, more correct first terms are 0.0473924,4 for $n = 5$ and 0.0551713,1 for $n = 10$.