Title: Some comments on the use of higher-order formula for numerical derivatives in scientific computing

Authors: N.Mohankumar[1,*] and Scott M. Auerbach [2,3]

Affiliation:

1. Radiological Safety Division
Indira Gandhi Centre for Atomic Research
Kalpakkam, India 603 102.
2. Department of Chemistry and
3. Department of Chemical Engineering, University of Massachusetts
Amherst, MA 01003 USA
***. Corresponding author**

Running title: Some comments on the use of higher-order formula for numerical derivatives in scientific computing

Address for correspondence:

N. Mohankumar
Radiological Safety Division
Indira Gandhi Centre for Atomic Research
Kalpakkam, India 603102
FAX: +91- 4114-27480235 (fax)
email: nmk@igcar.gov.in

Total number of pages of manuscript including abstract, title page and tables : 11
Number of tables: 4

Abstract

In an earlier article, we had indicated two applications where there was a significant improvement in accuracy due to the use of higher order approximation for certain spatial derivatives. The method of undetermined coefficients was used for this derivative approximation. In this note which is a sequel to the previous work, we provide a simple prescription for fixing the order of approximation of the required derivative. This procedure is based on the comparison of the various errors of the discretisation process.

PACS number: 02.60. Jh

Key words: Numerical differentiation, Method of Undetermined Coefficients, Radioactivity migration in Porous medium and MST formula

In a previous article [1], we had indicated an implementation of a higher order approximation for certain spatial derivatives which was based on the method of undetermined coefficients. This resulted in a significant improvement of accuracy for two specific problems. The first problem addressed the finite difference solution of a system of two coupled partial difference equations (pde). These pde's govern the migration of radioactivity in a porous medium [2]. The second problem pertains to the quantum mechanical evaluation of rate constant, using the Miller Schwartz Tromp (MST) formula [3]. This is for a bimolecular reaction involving an Eckart potential barrier [4,5]. In the first problem, the physical process of diffusion has to be assessed at two interfaces which is governed by a first-order derivative of concentration. These two interfaces are the waste matrix-fracture interface and the fracture-porous matrix interface. It is this first-order derivative term that was subjected to a higher-order approximation (a maximum of a *seven-point* approximation when saturation was reached). The forward time and centred space (FTCS) discretisation scheme was adopted for both the pdes's.

In the second problem a Discrete Variable Representation (DVR) Hamiltonian was chosen [6]. The required rate constant is expressed in terms of the derivatives of the eigen vectors evaluated at the point which separates the reactants and products. Here a first-order derivative is needed for the eigen vectors.

In this earlier article, no criterion was indicated for the choice of the order of approximation for the required derivative. We use as high an order as possible till numerical saturation sets in. Here we indicate a simple criterion which governs the order of approximation which is based on the comparison of various errors from the discretisation process. The pde for C_p , the concentration of radionuclide in the porous matrix is given by

$$\frac{\partial C_p}{\partial t} - \frac{D_p}{R_p}\frac{\partial^2 C_p}{\partial x^2} + \lambda C_p = 0, \quad b \leq x \leq B, \quad t \geq 0 \tag{1}$$

The pde for the concentration C, along the fracture is given by

$$\frac{\partial C}{\partial t} + \frac{\nu}{R}\frac{\partial C}{\partial z} - \frac{D}{R}\frac{\partial^2 C}{\partial z^2} + \lambda C + \frac{q}{Rb} = 0, \quad z \geq 0, \quad t \geq 0 \tag{2}$$

Here λ, R, D, ν and θ are the decay constant, the fracture retardation factor, the dispersion coefficient, the ground water velocity in the fracture and the matrix porosity respectively. In the last equation, the second, third and the fourth terms account for advection, diffusion and decay respectively. R_p and D_p are the retardation factor and the dispersion coefficient for the porous medium. Then q, the molecular diffusive flux crossing the fracture-porous matrix boundary is given by

$$q(z,t) \; = \; -\theta D_p \left\{ \frac{\partial C_p}{\partial x} \right\}_{x=b}, \quad z \geq 0, \quad t \geq 0 \tag{3}$$

The continuity of the particle current at the waste matrix-fracture interface under diffusion approximation is implied in the following equation.

$$-D \left\{ \frac{\partial C}{\partial z} \right\}_{z=0} + \nu C(0,t) \quad = \quad \nu C_o, \quad t \geq 0 \tag{4}$$

The rest of the initial and boundary conditions for solving these coupled pde's are given in our earlier paper and we do not repeat them for brevity. It must be emphasized that the concentrations $C(z,t)$ and $C_p(x,z,t)$ at every spatial point originates due to the diffusive fluxes at the wastematrix-fracture interface and the fracture-porous matrix interface. These diffusive fluxes in turn depend on the first-order derivatives in eqns.(3,4). Hence a higher order approximation is implemented precisely for these quantities through the method of undetermined coefficients described below.

The method of Undetermined Coefficients

The first derivative of $f(x)$ at the point x is approximated by the following $(m+n+1)$-point approximation [7,8].

$$\begin{aligned} f'(x) &= (1/h)[a_m f(x-mh) + a_{m-1} f(x-(m-1)h) + a_{m-2} f(x-(m-2)h) + ... \\ &+ a_1 f(x-h) + a_0 f(x) + b_1 f(x+h) + b_2 f(x+2h) + ... + b_n f(x+nh)] \end{aligned} \tag{5}$$

In this formula there are $(m+n+1)$ constants, $a_m, a_{m-1}, ..., a_0, b_1, b_2, ..., b_n$ which need to be fixed. For determining these constants, one can conveniently choose the point x as zero. Then by setting $f(x)$ to $x^0, x, x^2, x^3, ..., x^{m+n}$ successively and then equating the exact $f'(x)$ to the value of $f'(x)$ as determined by the above formula, one gets $(m+n+1)$ linear equations. These equations determine the $(m+n+1)$ constants. When $m = n$, this amounts to symmetric differencing. When $a_m, a_{m-1}, ..., a_1$ are all zero, this results in forward differencing. If $b_1, b_2, ..., b_n$ are all zero, then we have a backward difference approximation. For the porous flow problem, we need a forward difference scheme since the medium (i.e fracture) where the concentration is sought is defined only for $z \geq 0^+$. The minimum and the maximum orders of approximation that we have used are two and seven respectively. Here an n-point approximation is given by

$$\left\{\frac{\partial C}{\partial z}\right\}_{z=0^+} \simeq \frac{1}{\Delta z}\left\{a_0 C_{z=0^+} + b_1 C_{z=\Delta z} + b_2 C_{z=2\Delta z} + \ldots + b_{n-1} C_{z=(n-1)\Delta z}\right\} \tag{6}$$

A similar approximation is used for the derivative term $\frac{\partial C_p}{\partial x}$. Using the Mean Value Theorem, the error e_n of an n-point forward approximation is given by

$$e_n = \frac{1}{n}(\Delta z)^{n-1}\left\{\frac{\partial^n C}{\partial z^n}\right\}_{z=z_o}, \; z_o \in (0,(n-1)\Delta z). \tag{7}$$

Note that the validity of the above error formula demands the following conditions to be satisfied.

1. The existence of the derivatives $\frac{\partial^j C}{\partial z^j}, j = 0,1,2,..,(n-1)$ which are continuous in the closed interval $[0,(n-1)\Delta z]$.

2. The existence of the derivative $\frac{\partial^n C}{\partial z^n}$ in the open interval $(0, (n-1)\Delta z)$.

The existence of the partial differential eqns.(1,2) guarantees the existence of derivatives $\frac{\partial^j C}{\partial z^j}$ for $j = 0, 1$ which are continuous in the closed interval $[0, (n-1)\Delta z]$. In the remaining cases, we assume the existence of these required higher order derivatives.

Further for evolving a simple practical criterion, we tacitly assume that the derivative term in the error expression e_n is of the order of unity. We define E as

$$E = min\{E_1, E_2, ..., E_m\} \tag{8}$$

where $E_i's$ are the magnitudes of the leading discretisation errors of the remaining spatial derivative terms of the pde's. Then the required order of the derivative approximation is arrived at by demanding that

$$e_n \leq E \tag{9}$$

For convenience, we set $\Delta z = \Delta x = h$. The chosen values of the constants D_p, θ and b are $0.01, 0.01$ and 0.0005 respectively. The rest of the constant terms have values equal to unity. We need to consider the magnitudes of the leading truncation error for the following terms, $\frac{D_p}{R_p}\frac{\partial^2 C_p}{\partial x^2}$, $\frac{D}{R}\frac{\partial^2 C}{\partial z^2}$ and $\frac{\nu}{R}\frac{\partial C}{\partial z}$. Taking the derivative term in eqn.(7) as unity, the magnitude of the leading discretisation errors for these spatial derivative terms are $\frac{1}{100}\frac{h^2}{12}$, $\frac{h^2}{12}$ and $\frac{h^2}{6}$ respectively. The minimum of these three numbers is given by $E = \frac{1}{100}\frac{h^2}{12}$ which is the leading error of the derivative term $\frac{D_p}{R_p}\frac{\partial^2 C_p}{\partial x^2}$.

This quantity E must be the upper bound for the leading truncation error of the derivative approximant for

$$\frac{q}{Rb} = -\frac{\theta D_p}{Rb}\frac{\partial C_p}{\partial x} \tag{10}$$

which has the magnitude e_n given below.

$$e_n = \frac{\theta D_p}{Rb}\frac{h^{n-1}}{n} \tag{11}$$

Hence we have

$$e_n \leq E \Rightarrow \frac{h^{n-1}}{5n} \leq \frac{1}{100}\frac{h^2}{12} \Rightarrow \frac{h^{n-3}}{n} \leq \frac{1}{240} \tag{12}$$

With $h = 0.2$, to satisfy the above inequality the choice $n = 5$ is slighly little less than sufficient and for the same step size, the choice $n = 6$ is adequate. These conclusions are well supported by the results in table 1. Again, with the choice $h = 0.1$, a fifth-order formula should be good enough to satisfy the above inequality. This is also confirmed by the results of table 2. More importantly, we see the onset of essential saturation for $n = 7$ in the results of tables (1,2). This is well in conformity with our

simple analysis. The time steps for the calculations with step sizes $h = 0.1$ and $h = 0.2$ are kept at the same value to maintain identical errors stemming from the discretisation of the time derivatives in the pde's. It must be noted that as z increases, there is an over all deterioration of accuracy due to error accumulation. In tables 1 and 2, the number of converged digits indicates a definitive downward trend as z increases from $10m$ to $300m$.

Evaluation of the rate constant for the Eckart barrier

In this problem, a bimolecular chemical reaction of the type given below is considered.

$$A + BC \rightarrow AB + C \tag{13}$$

Here, one needs the rate constant k, for the formation of the molecule AB and it is defined by

$$\frac{d[AB]}{dt} = k[A][B] \tag{14}$$

Under certain assumptions, a quantum mechanically exact expression is provided by the Miller Schwartz Tromp [MST] formula [3]. Here, k is evaluated as a time integral of the flux-flux autocorrelation function $C_f(t)$.

$$kQ = \int_0^\infty dt C_f(t) \tag{15}$$

Here Q is the partition function for reactants. The Eckart potential barrier considered is of the type

$$V(s) = V_0\, Sech^2(s) \tag{16}$$

Here V_0 is a constant and s is the reaction coordinate. We construct a Hamiltonian matrix H in the Discrete Variable Representation (DVR) basis [6]. Diagonalization of H yields the eigenvalues $\{E_i\}$ and the eigenfunctions $\{\psi_i(s)\}$. The correlation function $C_f(t)$ is then given by [3]

$$C_f(t) = \sum_{i,j} exp\left[-\beta(E_i + E_j)/2\right] cos[(E_i - E_j)t/\hbar]\, \left|\langle i|\bar{F}|j\rangle\right|^2 \tag{17}$$

where

$$\left|\langle i|\bar{F}|j\rangle\right|^2 = (\hbar/2m)^2 |\psi_i'(0)\psi_j(0) - \psi_i(0)\psi_j'(0)|^2 \tag{18}$$

The derivative term of the eigen function ψ in the last equation is evaluated by the higher order approximation formula. The zero there defines the point (*the transition surface* in one space dimension) that separates the reactants and products.

There are few complications here. In the case of the porous medium, we could compare the truncation

errors of all the derivative terms which formed the basis of fixing the order of required approximation. In other words,the required order is fixed by demanding the leading discretisation error of this derivative term must be comparable to the truncation errors of the other derivative terms of the pde's. By analogy, we need to have an estimate of the reference error in the eigen values and eigen vectors as a function of N, the size of the basis functions. This error has contributions from various components. They stem from the approximation property of the basis functions, truncation of the number of basis function from infinity to a finite number and also from the diagonalization procedure employed. Unfortunately, precise estimates do not exist for the DVR basis. Numerical observations indicate that the accuracy lies some times close to an exponential order and it is mostly better than the order $O(1/N^2)$ [9,10]. We fix our attention on the quantity $C_f(0)$. The converged value of this quantity, accurate to the last digit evaluated here, by the choice $N = 300$ is 1.5012037.

Because the potential barrier is a symmetric function, we choose an *odd-order* approximation. Below, we list the 5,7,9 and 11 point approximations for the derivative and the associated leading truncation errors [7,8]. In the truncation errors below, we have omitted the derivative term as in the earlier case. Due to symmetry, we notice that the coefficient of the middle term, namely $\psi(0)$, is zero and so it does not figure in the formulae below.

$$\psi'(0) \simeq (1/h)\{(-1/12)[\psi(2h) - \psi(-2h)] + (2/3)[\psi(h) - \psi(-h)]\} \quad (19)$$

$$E = O(h^4/30) \quad (20)$$

$$\psi'(0) \simeq (1/h)\{(1/60)[\psi(3h) - \psi(-3h)] - (3/20)[\psi(2h) - \psi(-2h)] + (3/4)[\psi(h) - \psi(-h)]\} \quad (21)$$

$$E = O(h^6/140) \quad (22)$$

$$\psi'(0) \simeq (1/h)\{-(1/280)[\psi(4h) - \psi(-4h)] + (4/105)[\psi(3h) - \psi(-3h)]$$
$$- (1/5)[\psi(2h) - \psi(-2h)] + (4/5)[\psi(h) - \psi(-h)]\} \quad (23)$$

$$E = O(h^8/630) \quad (24)$$

$$\psi'(0) \simeq (1/h)\{(1/1260)[\psi(5h) - \psi(-5h)] - (5/504)[\psi(4h) - \psi(-4h)] + (5/84)[\psi(3h) - \psi(-3h)]$$
$$- (5/21)[\psi(2h) - \psi(-2h)] + (5/6)[\psi(h) - \psi(-h)]\} \quad (25)$$

$$E = O(h^{10}/2772) \quad (26)$$

Compared to the forward approximation of the porous flow problem, the error estimates here indicate a much better accuracy due to symmetry. With $h = 10/39$ which corresponds to $N = 40$, the values of E for the 5,7,9 and 11 point approximations are $1.44 \times 10^{-4}, 2.03 \times 10^{-6}, 2.97 \times 10^{-8}$ and 4.43×10^{-10}. In

order that these error estimates for the derivative to to be satisfied, the eigen vector ψ must be known at the discrete spatial points *exactly* within the machine precision. Alternatively, it is also sufficient if the precision of the function values is better than 4.43×10^{-10}, which is the maximum precision for the choice $n = 11$. However, ψ is not known exactly and the errors in ψ as a function of N are again not known precisely. Hence the convergence pattern displayed in table 3 does *not* conform to the calculated precision of the derivatives. Again, as a function of N, the convergence of the rseluts in table 3 is not exponential. The calculations were performed in $Matlab$ with a minimum of 14 digit precision.

Still, in table 3 we observe a very slow but definitive trend of improvement of accuracy as we increase n, the order of the derivative approximation. This is for the following reasons.

1. Due to symmetry there is a partial cancellation of systematic errors. That is the systematic error (not random error) in $\psi(jh)$ is partially cancelled by the systematic error of the term $\psi(-jh)$ since they occur in pair as $[\psi(jh) - \psi(-jh)]$. Here we assume that due to symmetry, the systematic errors in both $\psi(jh)$ and $\psi(-jh)$ have the same sign and perhaps they may have equal magnitude.

2. The derivative expression involves pairs of terms like $[\psi(jh) - \psi(-jh)]$. The multiplier coefficients for these pairs occur with *alternating* signs. Whenever the intrinsic sign of errors of two consecutive pairs is the same, partial error cancellation can still occur.

Finally, in table 4, we have presented the derivative of the first eigen vector at the origin for various values of N and n. Again, we do not see evidence of exponential convergence as a function of N. Even for $N = 300$, the maximum number of converged digits in the derivative (for a $11-$point approximation) is just six.

Conclusions

We have indicated a method of finding the order of approximation for some spatial derivatives which govern the accuracy critically for two physical problems. This is based on comparing the various errors arising out of the discretisation process. For the porous flow problem, the method predicts a satisfactory order beyond which saturation sets in. The method could not be applied for the Eckart barrier problem. This is not due to any intrinsic failure of the method but because the reference error estimates for the required quantities for this innately more complex problem do not exist. This we will address in future work.

Table 1. Concentration along the fracture as a function of z and n

delz=0.2; delx=0.2; dt= 0.004; B=100; z=500

z	Exact	n=2	n=3	n=4	n=5	n=6	n=7
10	0.62931	0.63334	0.62907	0.62919	0.62920	0.62920	0.62920
20	0.36274	0.37204	0.36264	0.36261	0.36263	0.36263	0.36263
30	0.18401	0.19572	0.18427	0.18402	0.184030	0.18403	0.18403
40	8.2194E-02	9.2464E-02	8.2539E-02	8.2180E-02	8.2175E-02	8.2176E-02	8.2176E-02
50	3.2303E-02	3.9364E-02	3.2640E-02	3.2324E-02	3.2313E-02	3.2313E-02	3.2313E-02
60	1.1193E-02	1.5161E-02	1.1425E-02	1.1219E-02	1.1208E-02	1.1208E-02	1.1208E-02
70	3.4266E-03	5.3036E-03	3.5508E-03	3.4446E-03	3.4372E-03	3.4368E-03	3.4368E-03
80	9.2898E-04	1.6916E-03	9.8306E-04	9.3800E-04	9.3433E-04	9.3405E-04	9.3404E-04
90	2.2357E-04	4.9374E-04	2.4327E-04	2.2719E-04	2.2571E-04	2.2557E-04	2.2556E-04
100	4.7879E-05	1.3234E-04	5.3989E-05	4.9083E-05	4.8584E-05	4.8533E-05	4.8528E-05
200		4.3037E-12	7.5596E-14	3.5903E-14	3.0974E-14	3.0055E-14	2.9867E-14
300		3.3128E-22	3.2637E-26	3.5583E-27	1.9720E-27	1.6742E-27	1.5982E-27

Table 2. Concentration along the fracture as a function of z and n

delz=0.1; delx=0.1; dt=0.004; B=100; z=500

z	Exact	n=2	n=3	n=4	n=5	n=6	n=7
10	0.62931	0.63128	0.62918	0.62922	0.62922	0.62922	0.62922
20	0.36274	0.36737	0.36263	0.36262	0.36263	0.36263	0.36263
30	0.18401	0.18991	0.18406	0.18400	0.18401	0.18401	0.18401
40	8.2194E-02	8.7305E-02	8.2228E-02	8.2137E-02	8.2137E-02	8.2137E-02	8.2137E-02
50	3.2303E-02	3.5774E-02	3.2360E-02	3.2277E-02	3.2276E-02	3.2276E-02	3.2276E-02
60	1.1193E-02	1.3105E-02	1.1239E-02	1.1185E-02	1.1183E-02	1.1183E-02	1.1183E-02
70	3.4266E-03	4.3054E-03	3.4533E-03	3.4248E-03	3.4238E-03	3.4238E-03	3.4238E-03
80	9.2898E-04	1.2728E-03	9.4110E-04	9.2895E-04	9.2844E-04	9.2842E-04	9.2842E-04
90	2.2357E-04	3.3966E-04	2.2810E-04	2.2375E-04	2.2354E-04	2.2353E-04	2.2353E-04
100	4.7879E-05	8.2081E-05	4.9306E-05	4.7979E-05	4.7908E-05	4.7904E-05	4.7904E-05
200		4.2272E-13	3.4520E-14	2.7019E-14	2.6337E-14	2.6264E-14	2.6256E-14
300		1.1583E-24	2.5828E-27	1.1385E-27	1.0105E-27	9.9241E-28	9.8961E-28

Table 3. Values of $C_f(0)$ as function of the size of the basis set N and for various orders of the derivative approximation n

N	$n=3$	$n=5$	$n=7$	$n=9$	$n=11$
32	1.1197581	1.4086619	1.4845197	1.5084213	1.5165466
34	1.1510977	1.4239183	1.4887871	1.5069842	1.5121748
36	1.1792996	1.4364824	1.4919758	1.5058309	1.5090691
38	1.2046683	1.4468257	1.4943450	1.5049008	1.5068546
40	1.2274916	1.4553461	1.4960995	1.5041521	1.5052734
42	1.2480391	1.4623755	1.4973966	1.5035521	1.5041432
44	1.2665583	1.4681882	1.4983553	1.5030732	1.5033344

Table 4. The derivative of the first eigen vector at origin for various values of N and n. All the entries in this table need a multiplier 10^{-10}.

N	n=5	n=7	n=9	n=11
40	0.24092001361190	0.31339519020627	0.30772057089543	0.30397992967817
80	0.49607311192342	0.50232136289964	0.50201289240796	0.50198399124454
200	0.48339809124954	0.48354139205768	0.48354022847642	0.48354039641536
300	0.58040190829375	0.58043529224745	0.58043497331359	0.58043486156445

References:

1. N. Mohankumar and S.M. Auerbach, Comput. Phys. Commun. , 161 (2004) 109.
2. C.T. Chen and S.H. Li, Waste Management, 17 (1997) 53.
3. W.H. Miller, S.D. Schwartz and J.W. Tromp, J. Chem. Phys., 79 (1983) 4889.
4. T.J. Park and J.C. Light, J. Chem. Phys., 85 (1986) 5870.
5. R.E. Wyatt, Chem. Phys. Letters, 121 (1985) 301.
6. D.T. Colbert and W.H. Miller, J. Chem. Phys., 96 (1992) 1982.
7. H.M. Antia,“ Numerical Methods for Scientists and Engineers ”, Tata McGraw-Hill Book Company, New Delhi, 1991, pp. 160.
8. C. Pozrikidis,“ Numerical Computation in Science and Engineering ”, Oxford University Press, New York, 1998, pp. 319.
9. H. Wei, J. Chem. Phys., 106 (1997) 6885.
10. R.G. Littlejohn, M. Cargo, T. Carrington, Jr and B. Poirier, J. Chem. Phys., 116 (2002) 8691.