JOURNAL OF COMPUTATIONAL PHYSICS 28, 426–432 (1978)

Note

A Note on the Roundoff Error in the Numerov Algorithm*

Roundoff error is investigated in solutions of the partial wave Schroedinger's equation by Numerov's method. The uncommon summed form of the algorithm is shown to be superior in this regard and is recommended. Various details and analytic solutions are discussed.

Since its origin in solar system astronomy, Numerov's algorithm [1, 2] for the numerical solution of differential equations (DEQ) of the type $d^2y/dx^2 = A(x)y(x)$ has found a permanent home in physics: in optics and the physics of atoms, molecules, nuclei, fluids, and solids. Although many have stated [3–5] that it is the "best" method for solving such equations, a variety of competing methods exist of both the "shooting" (step by step) [6, 7, 8] and matrix [9] types. While we don't wish to enter the controversy over which method is best, if indeed one is best, we do wish to respond to some recent [8] criticism of the Numerov method and to restate some lessons which seem to have been forgotten in the ten years since Ref. [3] was published. The latter article contains a wide variety of practical information on the solution of the continuum Schroedinger equation and is highly recommended. Because scattering data have recently become available at high energies and at all angles, the need for highly accurate phase shift codes has dramatically increased [10, 11]. Accurate codes are also necessary in order to test approximation schemes, such as the eikonal approximation [12]. Our comments are relevant to these needs also.

Sources of numerical error in solving DEQ's by shooting methods are of three basic types: truncation error caused by mapping the DEQ into an "equivalent" problem on a finite mesh, roundoff error produced by the finite word size of computers, and matching errors caused by whatever scheme is used to extract useful information (phase shifts, binding energies, etc.) from the solution y. We are interested here primarily in the roundoff errors. In a recent comparison of binding energies calculated using a perturbative method on the one hand and the Numerov algorithm on the other hand, it was found that the latter method produced large roundoff errors and was rather strongly criticised for this. We wish to point out that there exist several forms of the Numerov algorithm, one of which is the ordinary form developed by Numerov, described in standard texts [13] and commonly used. Another is the summed form of the algorithm[3] which has an identical truncation error, is *extremely stable* against certain types of roundoff error, and effectively gives one the first derivative of y,

* Work performed under the auspices of the U. S. Energy Research and Development Administration.

0021-9991/78/0283–0426$02.00/0

as well. Although summed forms of DEQ algorithms are of considerable antiquity, there has been relatively little discussion of them [3, 14]. We are aware of only two explicit references [15, 16] to the use of this form of the Numerov algorithm since the discussion in Ref. [3]. Indeed, roundoff error itself is usually not extensively discussed (see, however, Refs. [13, 14, 17, 18]).

We wish to make a comparison of the ordinary and summed forms of the algorithm for a particularly simple problem where the exact solution is known. We will compute partial wave phase shifts in the absence of a potential; the exact result is zero, but errors in the numerical calculation will produce a non-zero result. Since one finds *a posteriori* that there are no qualitative differences between results for various partial waves, we will present results for the lowest partial wave, $l = 0$, where the Numerov *difference* equations (including roundoff error) have exact solutions, which also are instructive. The other partial waves require a more detailed treatment following the methods of Ref. [3] or [19], but do not generate additional insight. The error analysis for bound states is virtually identical.

The ordinary Numerov algorithm consists of a single difference equation defined on an equally spaced mesh, $x_n = nh$, where h is the spacing. Defining $\xi_n = (1 - h^2A_n/12)y_n$ and $T_n = h^2A_n/(1 - h^2A_n/12)$ with $A_n = A(x_n)$, $y_n = y(x_n)$ we have

$$\xi_{n+1} = (2 + T_n)\,\xi_n - \xi_{n-1} \tag{1a}$$

if we neglect roundoff error. Roundoff error occurs when the right-hand side of Eq. (1) is calculated using the appropriate values stored in the computer and the result is actually given by RHS $\cdot\,(1 - \epsilon_n)$, where RHS is the *exact* value one would have obtained with infinite word size. Rather than carry out a statistical analysis [14, 17, 18] we simply replace the fractional error at the nth step, ϵ_n, by ϵ, the average value (which will be positive) and solve the difference equation which actually results from the computer's operation

$$\xi_{n+1} = ((2 + T_n)\,\xi_n - \xi_{n-1})(1 - \epsilon). \tag{1b}$$

For the reduced Schroedinger equation with no potential and $l = 0$, we find $A_n = -k^2$, where $E = k^2/2m$ is the energy of a particle of mass m. Since T_n is also a constant T, Eq. (1b) can be solved exactly [20] for $n \geqslant 1$ subject to the usual boundary conditions: $\xi_0 = 0$, $\xi_1 =$ constant. With $\lambda = 1 - \epsilon + (T/2)(1 - \epsilon)$ we find

$$y_n = B\beta^n \sin(n\phi) \cong B(1 - x_n\epsilon/2h)\left[\sin(kx_n) + kx_n\left(\frac{(kh)^4}{480} + \frac{\epsilon}{2(kh)^2}\right)\cos(kx_n)\right], \tag{2a}$$

$$\beta = (1 - \epsilon)^{1/2} \cong 1 - \epsilon/2 \qquad \phi = \cos^{-1}(\lambda/\beta) \cong kh + (kh)^5/480 + \epsilon/(2kh), \tag{2b}$$

where the second form in Eq. (2a) follows from the approximate forms of Eq. (2b). By either matching y_n at two adjacent points [3] to the linear combination $B(\sin(kx_n) +$

$\tan(\delta_0)\cos(kx_n))$ or by matching derivatives in the usual way, the phase shift δ_0 may be determined to leading order in terms of $R \equiv kx_n$

$$\delta_0 \cong (R - \sin R \cos R)\left(\frac{\epsilon}{2(kh)^2} + \frac{(kh)^4}{480}\right). \tag{3}$$

The expression for δ_0 displays the characteristic h^4-dependence of the algorithm's truncation error and the $1/h^2$ roundoff error [3.19]. Note that it has a minimum at $kh = (120\epsilon)^{1/6}$ for fixed R.

The summed form of the algorithm is derived in Ref. [14, 18]. Basically one writes n successive elements of the difference equation and adds them. *Including roundoff error*, the summed algorithm has the form

$$F_n = (F_{n-1} + T_n\xi_n)(1 - \epsilon_1), \tag{4a}$$

$$\xi_n = (\xi_{n-1} + F_{n-1})(1 - \epsilon_2) \tag{4b}$$

subject to $\xi_0 = 0$, $\xi_1 = \text{const.}$ for $n \geqslant 1$. In addition, ϵ_1 and ϵ_2 are the separate averaged roundoff errors for Eqs. (4a) and (4b), ξ_n is the same as before, and $F(x_n)$ is $(hy'(x_n) + \text{order } (h^2))$. One may evaluate $y'(x_n)$ up to (but not including) order (h^6) using the expressions

$$hy'_n = \alpha \sum_{m=-3}^{2} F_{(m)} + \beta \sum_{m=-2}^{1} F_{(m)} + \gamma \sum_{m=-1}^{0} F_{(m)}\,, \tag{5a}$$

where

$$F_{(m)} \equiv F_{n+m} \tag{5b}$$

and

$$\begin{aligned} \alpha &= 7/720, \\ \beta &= -29/360, \\ \gamma &= 91/144. \end{aligned} \tag{5c}$$

Thus, although the assertion of Ref. 11 and 21 is correct that the Numerov method (in its ordinary form) does not generate the derivative of y, this is not true for the summed form and allows calculation of derivatives *without differencing*, a decided advantage. For the model problem we solved earlier we may exactly solve for either y_n or F_n in the same way. Defining $\epsilon' = (\epsilon_1 + \epsilon_2)$, $\lambda' = 1 - \epsilon'/2 + T(1 - \epsilon')/2$ and $\beta' = \sqrt{1 - \epsilon'}$ we find

$$y = B\beta'^n \sin(n\phi_0) \cong B(1 - R(\epsilon'/2kh))(\sin(R) + (R(kh)^4/480)\cos(R)) \tag{6a}$$

$$\phi_0 = \cos^{-1}(\lambda'/\beta') \cong kh + (kh)^5/480 + \text{order}(\epsilon' h), \tag{6b}$$

$$F_n = \xi_1\beta'^n(\cos(n\phi_0) + \alpha' \sin(n\phi_0)) \cong \xi_1(1 - R(\epsilon'/2kh)) \times \left(\cos(R) - \left[\frac{\epsilon_1 - \epsilon_2}{2kh} + \frac{(kh)^4 R}{480} + \frac{kh}{2} + \frac{(kh)^3}{24}\right] \sin(R)\right), \tag{6c}$$

where α' was adjusted to generate the correct F_1. Obtaining y' from Eq. (5) (δ_s') or calculating y' directly from Eq. (6a) (δ_s), the phase shifts may be calculated with *different* results.

$$\delta_s \cong \frac{\epsilon'}{2kh} \sin^2 R + (kh)^4(R - \sin R \cos R)/480, \tag{7a}$$

$$\delta_s' \cong \frac{(\epsilon_1 - \epsilon_2)}{2kh} \sin^2 R + (kh)^4(R - \sin R \cos R)/480. \tag{7b}$$

Note that, unlike the $(1/h^2)$-behavior of δ_0, δ_s behaves like $(1/h)$. In addition we would expect that $\epsilon_2 \cong \epsilon_1$ and therefore that $\delta_s' \cong 0$ for small h.

We calculated (for $R = 20.0$) the first ten phase shifts using both the ordinary and summed algorithms for a variety of starting constants, ξ_1, for each value of kh. For the summed algorithm Eq. (5) was used to calculate y'. The starting values affect results only by generating different roundoff errors. For a given value of kh, means and variances were calculated for each set of starting values. In addition, predictions were calculated using Eq. (3) and (7b), respectively. This requires a value for ϵ, which depends on the coding. A particularly simple analysis and minimal roundoff is possible if one calculates Eq. (1a) by first forming $(2\xi_n - \xi_{n-1})$ and then adding $T_n\xi_n$. There is no error *produced* by the first operation and the second produces a single *chopping*-type roundoff error [3]. For a machine with a floating point mantissa of m bits, we assume that the non-significant bits in any operation are randomly distributed from zero to a maximum value 2^{-m+1} and a mean value $\epsilon_0 = 2^{-m}$. The significant bits are *not* randomly distributed (the first must be a one) and we assume they follow the 1/z-distribution discussed by Hamming [22]. This leads immediately to $\epsilon_1 = \epsilon_2 = \epsilon = \epsilon_0/2\ln(2) \cong 0.721\ \epsilon_0$. Calculations were performed on a CDC 7600 which has $m = 47$.

The results are plotted in Figs. 1 and 2. The individual points are calculated using the algorithm and the dashed, dashed-dot, and solid lines are the separately calculated roundoff, truncation, and total errors, respectively, from Eqs. (3) and (7b). In Fig. 2 the roundoff in the total was assumed to be zero ($\epsilon_1 = \epsilon_2$). For comparison the dashed line includes roundoff assuming $\epsilon_1 - \epsilon_2 = \epsilon$. From the results it is clear that $\epsilon_1 \cong \epsilon_2$. The difference between the two figures is very dramatic and illustrates the superiority of the summed algorithm. Both calculations take the same time and storage. In many applications there will be no advantage at all to using the ordinary algorithm, and we therefore strongly suggest that the summed form be used. The predicted errors are also in excellent agreement with actual errors and the ϵ we actually find agrees with our prediction to considerably better than one percent.

Our results also indicate that using Eq. (5) to calculate logarithmic derivatives will be superior to differencing the y's. Both roundoff and matching errors will be smaller. This applies to both phase shift and bound state calculations; both use the logarithmic derivative and both produce similar errors.

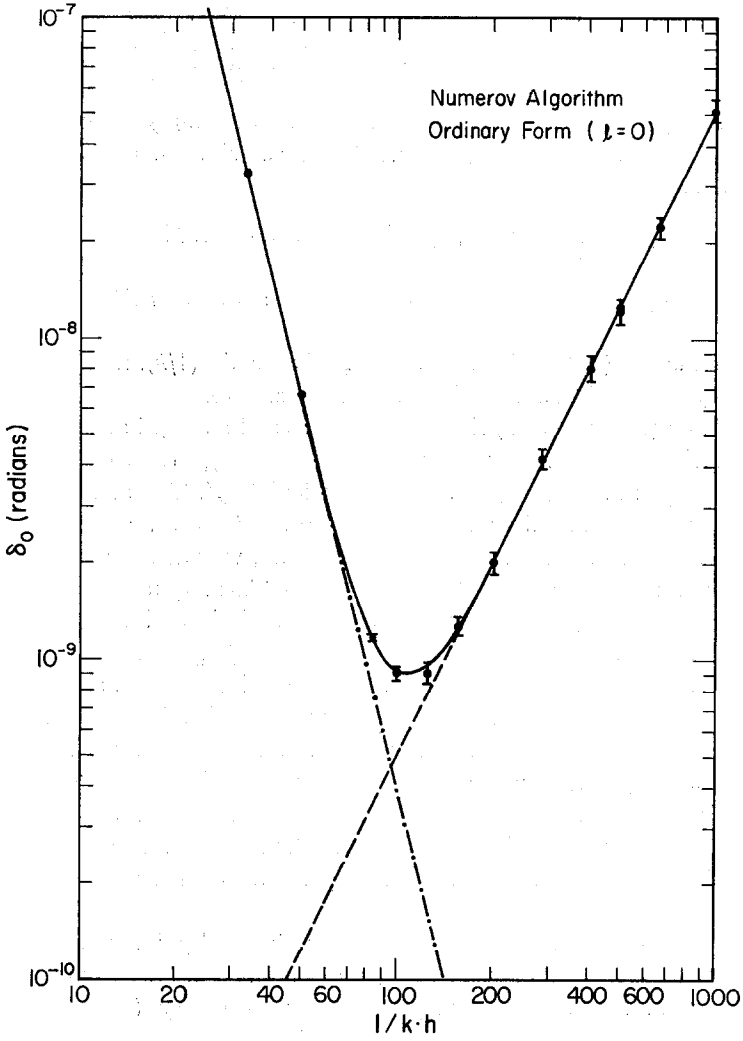

FIG. 1. Phase shifts for $l = 0$ calculated using the ordinary form of the Numerov algorithm. Calculated points are indicated by dots, while predictions for roundoff, truncation, and total errors are indicated by dashed, dashed-dot, and solid lines.

Finally we discuss the results of Ref. [8]. Although they do not state which version of the algorithm they used, it is clear that the eigenvalues in Table VI of that work have the $(1/h^2)$-roundoff error typical of the ordinary form of the algorithm. While we cannot predict that the summed form would produce results as stable as the preferred method of Ref. [8], it is abundantly clear that they would be far superior to the Numerov results displayed in Ref. [8]. In all fairness this is the comparison which should be made when roundoff error is a significant factor.

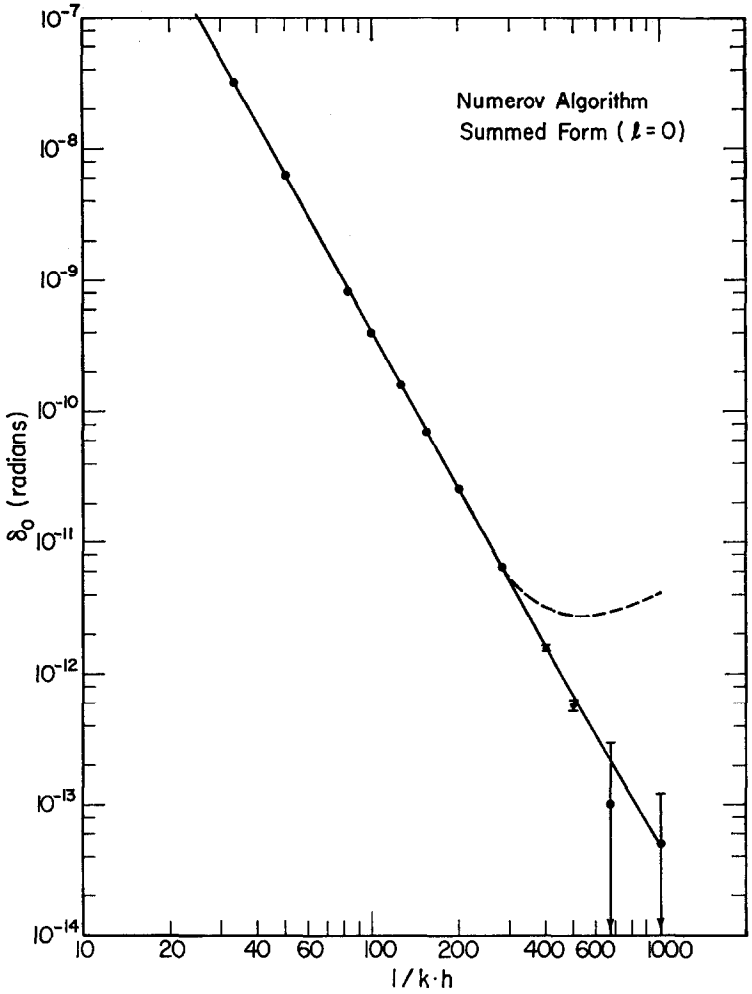

FIG. 2. Same as in Fig. 1 for the summed algorithm. Solid line contains no roundoff error while the dashed line does, as discussed in the text.

REFERENCES

1. B. V. NOUMEROV, *Publ. Observ. Cent. Astrophys. Russ.* **2** (1923), 188.
2. B. V. NOUMEROV, *Mon. Notic. Roy. Astron. Soc.* **84** (1924), 180; (1924), 592.
3. M. A. MELKANOFF, T. SAWADA, AND J. RAYNAL, *Methods Comp. Phys.* **6** (1966), 1.
4. J. M. BLATT, *J. Computational Physics* **1** (1967), 382.
5. C. FROESE, *Can. J. Phys.* **41** (1963), 1895.
6. F. Y. HAJJ, H. KOBEISSE, AND N. R. NASSIF, *J. Computational Physics* **16** (1974), 150.
7. A. C. ALLISON, *J. Computational Physics* **6** (1970), 378.
8. GH. ADAM, L. GR. IXARU, AND A. CORCIOVEI, *J. Computational Physics* **22** (1976), 1.
9. B. W. SHORE, *J. Chem. Phys.* **59** (1973), 6450.

10. R. R. DOERING, A. I. GALONSKY, AND R. A. HINRICHS, *J. Computational Physics.* **12** (1973), 498.
11. R. A. EISENSTEIN AND G. A. MILLER, *Comp. Phys. Comm.* **8** (1974), 130.
12. S. J. WALLACE, *Ann. Phys. (New York)* **78** (1973), 190.
13. R. W. HAMMING, "Numerical Methods for Scientists and Engineers," McGraw–Hill, New York, 1962.
14. P. HENRICI, "Discrete Variable Methods in Ordinary Differential Equations," pp. 327–343, Wiley, New York, 1962.
15. A. C. YATES, *Phys. Rev.* **176** (1968), 173.
16. A. CHUTJIAN, *J. Chem. Phys.* **51** (1969), 5414.
17. P. HENRICI, "Elements of Numerical Analysis," p. 302, Wiley, New York, 1964.
18. P. HENRICI, in "Proceedings of a Symposium on the Numerical Treatment of Ordinary Differential Equations, Integral Equations, and Integro-Differential Equations, Rome, 1961."
19. J. N. BASS, *J. Computational Physics* **9** (1972), 555.
20. F. B. HILDEBRAND, "Methods of Applied Mathematics," Prentice–Hall, Englewood Cliffs, N.J., 1952.
21. J. W. COOLEY, *Math. Comp.* **15** (1961), 363.
22. Ref. [13], p. 37, and the references cited there.

RECEIVED: June 30, 1977; REVISED: October 19, 1977

J. L. FRIAR

Theoretical Division
Los Alamos Scientific Laboratory
University of California
Los Alamos, New Mexico 87545