Response to “Convergence of Summation-by-Parts Finite Difference Methods for the Wave Equation”

Svärd Magnus and Nordström Jan

The self-archived version of this journal article is available at Linköping University Electronic Press:

<http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-138577>

N.B.: When citing this work, cite the original publication.

The original publication is available at www.springerlink.com:

Magnus, S., Jan, N., (2017), Response to “Convergence of Summation-by-Parts Finite Difference Methods for the Wave Equation”, *Journal of Scientific Computing*, 218. <https://dx.doi.org/10.1007/s10915-017-0468-x>

Original publication available at:

<https://dx.doi.org/10.1007/s10915-017-0468-x>

Copyright: Springer Verlag (Germany)

<http://www.springerlink.com/?MUD=MP>

RESPONSE TO "CONVERGENCE OF SUMMATION-BY-PARTS FINITE DIFFERENCE METHODS FOR THE WAVE EQUATION"

MAGNUS SVÄRD AND JAN NORDSTRÖM

ABSTRACT. This is a short response to some statements by Wang and Kreiss in "Convergence of Summation-by-Parts Finite Difference Methods for the Wave equation" ([WK16]) that questions results in our paper "On the order of accuracy for difference approximations of initial-boundary value problems" ([SN06]). We show that our results still stand.

1. THE PROBLEM AND EARLIER RESULTS

The article [WK16] analyses convergence rates of finite difference schemes for the wave equation in one dimension with Neumann boundary conditions. This has to some extent already been done in [SN06].

Briefly summarised: in [SN06] we considered schemes approximating initial-value problems and assumed that they were L^2 -stable (i.e., energy stable) and that the solution was bounded uniformly at every point in space and time. If so, we showed that equations with a p th order derivative in space and a boundary truncation error of order q (including all terms and boundary conditions), will have a convergence rate of $p + q$, or the internal order of accuracy whichever is smaller. These results rely on *initial data and boundary data being compatible and smooth, and the forcing function being smooth*. Furthermore, the scheme has to be *consistent* with the partial differential equation (PDE).

2. CLAIMS BY WANG AND KREISS

We list the claims in [WK16] that challenges results in [SN06].

- (1) Wang and Kreiss refer to our results as "optimal" convergence rates.
- (2) In Section 5.2 they present a numerical example that allegedly disproves certain results in [SN06].
- (3) They suggest that the singularity of a certain boundary matrix contradicts our previous results. (See the last paragraph of Appendix 2, [WK16].)

The last two statements questions at least part of the main result in [SN06].

3. RESPONSES

Claim 1: We never once claim in [SN06] that our results were *optimal*. We are fully aware that more detailed information can be obtained from Laplace transforming a specific problem as is done in [WK16].

Furthermore, we emphasize that the purpose of [SN06] was not optimality, but to provide *relatively simple conditions* for the users of finite differences to check. In contrast to this, the Laplace transform analysis can be rather cumbersome.

Claim 2: The numerical example in Section 5.2 in [WK16] solves the following PDE.

$$(1) \quad e_{tt} = e_{xx} + F(x, t), \quad 0 < x < 1$$

with Neumann boundary conditions, for the exact solution, $e_{ex} = \cos(10\pi x + 1)\cos(10\pi t + 2)$, $0 \leq x \leq 1$ with $F \equiv 0$. The following stable and consistent scheme is used to solve (1)

$$(2) \quad \mathbf{e}_{tt} = -H^{-1}A\mathbf{e} + \mathbf{F},$$

where $\mathbf{e}(t)$ is the solution vector. (The difference operator is $D_2 = H^{-1}(-A + BS)$. For further details and notation, we refer to [WK16].) They further state that F is $(10\pi)^3 \approx 3.1 \cdot 10^4$ at the boundary point and 0 elsewhere. It is unclear if they refer to F in (1) or $\mathbf{F}$ in (2) or both of the previous. Either way, since initial and boundary data are given by the exact solution e_{ex} , they must expect the discrete solution to converge to e_{ex} . (And it will, since the solution is the same in the sense that $\|e_{F \neq 0} - e_{ex}\|_2 \rightarrow 0$, $h \rightarrow 0$. It does not matter which of the three interpretations that was intended.) They obtain an error that decreases as $\mathcal{O}(h)$, which they claim contradicts [SN06] which they in turn claim predicts a $\mathcal{O}(h^2)$ convergence rate. Clearly, one can not expect second-order convergence since [SN06] require smooth data, including the forcing function which is violated in all 3 options.

This should have been the end of this discussion. However, the two referees of this note both claim that the example actually is to be understood as the error equation and hence that $\mathbf{F}$ is the truncation error with initial and boundary data being zero. We remark that this is never stated in [WK16]. Nevertheless, we welcome this claim since it gives us the opportunity to clarify the relation between a scheme and its error equation. We will address this claim by making three observations.

Firstly, equation (1) discretized by (2) is a completely general problem, and there is nothing intrinsic to this problem that makes it solely an error equation. To see this, consider the following function, $u = e + \Phi$ where $\Phi(x, t)$ is a smooth function. By choosing Φ appropriately, we can construct the following problem,

$$(3) \quad u_{tt} = u_{xx} + G(x, t)$$

with arbitrary non-homogeneous smooth boundary and initial data, and $G = F - \Phi_{xx} + \Phi_{tt}$. This is simply a transformation of the original problem with completely general data and an arbitrarily chosen discontinuity in the first element of the forcing. Since it is merely a transformation, it retains the same convergence properties. Clearly, the general problem (3) does not satisfy the smoothness requirements of [SN06], since G is not smooth. Hence, [SN06] does not predict second-order convergence rate of (3), let alone the special case (1).

Secondly, we note that if (2) is an error equation, then the governing PDE is

$$(4) \quad \begin{aligned} u_{tt} &= u_{xx} + R(x, t), \quad 0 < x < 1 \\ u_x(0, t) &= u_x(1, t) = 0 \end{aligned}$$

with bounded initial data and where R is a smooth forcing function. It is discretized by means of

$$(5) \quad \mathbf{v}_{tt} = D_2\mathbf{v} - H^{-1}(BS\mathbf{v} - 0) = H^{-1}(-A + BS)\mathbf{v} - H^{-1}(BS\mathbf{v} - 0) + \mathbf{R}.$$

or equivalently

$$(6) \quad \mathbf{v}_{tt} = -A\mathbf{v} + \mathbf{R}.$$

Inserting the exact solution into (5) or (6) and subtracting the semi-discrete solution results in the error equation (2) where $\mathbf{F}$ is the truncation error. However, with the scheme (5) or (6), $\mathbf{F}$ would be $\mathcal{O}(h)$ at the boundary, not $\mathcal{O}(1)$. If $\mathbf{F}$ is $\mathcal{O}(1)$,

our second observation is that this would imply that the scheme is inconsistent and hence not obeying the conditions of [SN06].

Thirdly, it may be tempting to think that one could add a slightly smaller error (say $\mathcal{O}(h^{1/2})$) and thereby making the scheme (kind of) consistent. To preempt the idea that this is a viable test, we proceed to our third observation. In order for (2) to be a valid “model” for a true error equation, the modelled truncation error must be possible to obtain without altering the properties in [WK16] that led to their conclusions regarding the “gain of accuracy”. Hence, at the core of this problem is the question: Is it possible to change the scheme such that the truncation error becomes $\mathcal{O}(h^\alpha)$, $0 \leq \alpha < 1$, while retaining the properties derived in [WK16]?

To maintain the convergence properties, A can not be changed. (Any change in A would change the eigenstructure of the boundary matrix and thereby the analysis in [WK16].) Hence, we are forced to modify BS in order to lower the boundary accuracy. In view of (5), any error added to the boundary derivative will cancel against the same term, but with opposite sign appearing in the penalty. This is obvious in (6) where BS does not appear. Finally, and to be exquisitely clear, we can not change BS in only one of the two places where it occurs, since then the eigenstructure would change (effectively changing A), and the energy estimate would be destroyed. As a consequence, *the numerical example has no connection to the problem analyzed in [WK16]*. The example stands by itself and represents nothing but a problem with a discontinuous forcing. It is not a counterexample of [SN06].

Claim 3: This statement concerns a certain boundary matrix that appears in the Laplace transform analysis. See e.g., equation (37) and the first equation in Section 3.2.2 in [WK16] and equation (12) in [SN06]. The boundary matrix being singular does not contradict the results in [SN06].

Firstly, in the proof of Theorem 2.6, the inverse of the boundary matrix is not used. It is a misconception that it has to be non-singular. It is only if the goal is to prove stability (and the method is not *a priori* known to be stable), that the non-singularity is a necessary condition. In [SN06], we turn things around and consider *a priori stable methods* and use their stability properties to infer convergence rates via the boundary matrix.

Secondly, the analysis in [WK16], shows that a “gain of accuracy” is obtained despite the matrix being singular. Indeed, this shows that the singularity is not an issue.

Thirdly, the sole reason why the singularity of the boundary matrix is discussed, is that it was offered as an explanation for the “suboptimal” convergence results in the numerical example (Section 5.2, [WK16]) discussed at length here. Since this example does not show “suboptimal” convergence, the “singularity explanation” is clearly misplaced.

4. AN ILLUMINATING EXAMPLE

Since the numerical example that supposedly corroborated the theoretical results in [WK16] turns out not to do that, there is really no reason to claim that their theoretical result (one order gain at the boundary) derived in Section 3.2.1 is sharp. In order to test its sharpness, one could try to devise a numerical example that would emphasize the boundary errors, rather than the internal errors. (This is not easy as observed in [WK16].) To this end, we will run the the second-order scheme (2), i.e., *exactly the same scheme as in Section 5.2, [WK16], but with different initial conditions and a different (smooth) forcing function*.

N	L^2 -error	rate	L^∞ -error	rate
5	9.20e-3	-	1.75e-2	-
10	1.69e-3	2.44	3.08e-3	2.50
15	6.96e-4	2.19	1.34e-3	2.04
20	3.77e-4	2.13	7.21e-4	2.16
25	2.36e-4	2.09	4.47e-4	2.14

TABLE 1. Convergence rates for second-order scheme at $T = 0.4$.

To highlight the boundary errors, we choose $u(x, t) = \cos(t)(x^2 - \frac{2}{3}x^3)$, on the domain $0 < x < 1$. It satisfies $u_x = 0$ on both boundaries. We use this function to obtain initial conditions and derive a forcing function such that u is the solution of $u_{tt} = u_{xx} + F(x, t)$. The function u has several properties that makes it suitable for this test. 1) A third-order polynomial is exactly differentiated by the interior scheme. This will minimize the error contribution from the interior. 2) The boundary schemes will not differentiate this function exactly. Together with the first property, boundary errors should dominate the global error when there are few grid points. (With more grid points the internal error, fed by the boundary error, will dominate.)

Furthermore, the Taylor series of the boundary error will only contain the leading error term. Recall that generally, $error = c_1 h^p + c_2 h^{p+1} + \dots$ and only when h is small will we see the rate p . However, in this case only $c_1 \neq 0$ and hence we should be able to register the rate p even for large h . Our choice of solution is tailored to measure boundary errors.

We ran the scheme (2) (the second-order version with first-order boundary closures) until $T = 0.4$ and measured the errors. The time step was very small, $\Delta t = 0.05 \cdot \Delta x$. The standard 4th-order Runge-Kutta scheme was used to march in time. The results are presented in Table 1. (N denotes the number of grid points.)

The L^2 convergence rate is greater than 2 on course meshes, and as the number of interior points increases, the rate drops towards 2. (Similar results are obtained at other times as well.) The higher than 2 convergence rates, indicates that the boundary errors are smaller than the internal errors.

Furthermore, the recorded L^∞ errors also shed light on this problem. The reason why there is a “gain of accuracy” in the L^2 -norm of the boundary error, is that the error quickly drops off into the domain. Assuming that the range of the boundary error is proportional to the grid size, we should see a convergence rate in L^∞ of 2.5, if the L^2 contribution of the boundary error is $\mathcal{O}(h^3)$, i.e., if two orders are gained. (To see this we assume that we have fixed number of points, say M , with error h^p . The L^2 contribution is then, $\|e_b\|_2^2 = hM(h^p)^2$. If two orders are gained $\|e_b\|_2 \sim \mathcal{O}(h^3)$ and $p = 2.5$.) If one order was gained an L^∞ rate of 1.5 would be expected when the boundary errors dominate (small N). Note that even if we had chosen another range of influence of the boundary error, the result would be higher than 2 in L^∞ if two orders are gained, and less than 2 if one order is gained. It is fairly clear which one of these two hypotheses that are supported by the results in Table 1.

These computations, although not conclusive, indicate that the analysis in Section 3.2.1 [WK16] may not be sharp.

5. FINAL REMARKS

The authors of [WK16] fail to mention that in Theorem 1, the accuracy in the marginal case for stability drops below the expected two orders. This may at first

glance appear as a “counterexample” but incidentally, that is also the exact point where the solution ceases to be in L^∞ . (In the marginal case, $\tilde{A}$ in (9) has a rank deficit and can not bound the solution in H^1 .) This is in perfect agreement with [SN06].

Taking a critical view of our paper [SN06], we note that we occasionally refer to the boundary matrix as being non-singular or invertible in some of the latter proofs. Since that is not required in Theorem 2.6, on which the remaining theorems are based, this is a careless wording and we should have written that the “ σ -vector is uniquely determined”.

The singular case that results from the wave equation with Neumann boundary conditions is interesting and we will in an upcoming paper provide further insights into the properties of the boundary matrix. However, our main message here is that the analysis and results in [WK16], although interesting in their own right, do not contradict [SN06].

REFERENCES

- [SN06] M. Svård and J. Nordström. On the order of accuracy for difference approximations of initial-boundary value problems. *J. Comput. Physics*, 218(1):333–352, October 2006.
- [WK16] S. Wang and G. Kreiss. Convergence of summation-by-parts finite difference methods for the wave equation. *J. Sci. Comput.*, in press, 2016.

DEPT. OF MATHEMATICS, UNIVERSITY OF BERGEN, P.O. BOX 7803, 5020 BERGEN, NORWAY
E-mail address: `Magnus.Svard@uib.no`

COMPUTATIONAL MATHEMATICS, DEPARTMENT OF MATHEMATICS, LINKÖPING UNIVERSITY, SE-581 83 LINKÖPING, SWEDEN
E-mail address: `jan.nordstrom@liu.no`