A Correspondence between Errors and Pseudo-errors of Approximate Computations with Similar Rates of Convergence

Aram Soroushian

Structural Engineering Research Centre, International Institute of Earthquake Engineering and Seismology, No. 21, West Arghavan, North Dibajee, S. Lavasani (Farmaniyeh), Tehran 19537, Iran

a.soroushian@iiees.ac.ir, aramsoro@yahoo.com

Abstract

A major requirement of arbitrary approximate computation is convergence to the exact solution. In view of the definition of error, problems with analytical closed from solutions generally need to be at hand, for numerical study of convergence, as well as verification and validation issues. In order to eliminate this necessity, the concepts of pseudo-error and pseudo-convergence are introduced in 2010. Towards better control of accuracy, it is herein explained that, if two solutions, obtained from two different methods, for one problem, converge to the exact solution with similar rates, and in intervals of proper convergence, and the errors of the first solution are more/less than the errors of the second solution, the same can be claimed about the two pseudo-errors, and vice versa. Extension to several computations is concluded and the implementation in practice is briefly discussed.

Keywords: Convergence, Exact solution, Error, Pseudo-error, Rate of convergence, Proper convergence

Introduction

Convergence is the most important essentiality of arbitrary approximate computations [1, 2]. With the purpose to study convergence and its trend numerically and without the exact solutions, the concepts of pseudo-error and pseudo-convergence are introduced and later extended to non-geometric changes of the algorithmic parameters [3-5]. In brief, for an arbitrary approximate computation $\mathbf{U}^a$ defined in terms of the algorithmic parameter λ, by defining the pseudo-error D, as (q stands for the rate of convergence [1-5], the right subscript i implies the result of the computation when using λ_i as the value of the algorithmic parameter, and $\| \ \|$ represents an arbitrary norm [6]):

$$D_i = \frac{\left\| \mathbf{U}^a{}_i - \mathbf{U}^a{}_{i-1} \right\|}{\left(\frac{\lambda_{i-1}}{\lambda_i} \right)^q - 1} \quad , \qquad \lambda_i < \lambda_{i-1} \tag{1}$$

the equivalence addressed in Fig. 1 holds, i.e. either both the errors, E, and the pseudo-errors, D, or neither of them, converge properly. Proper convergence [7] is defined as the decrease of the errors in the log-log convergence plot as a straight line sloped q (q is defined just before

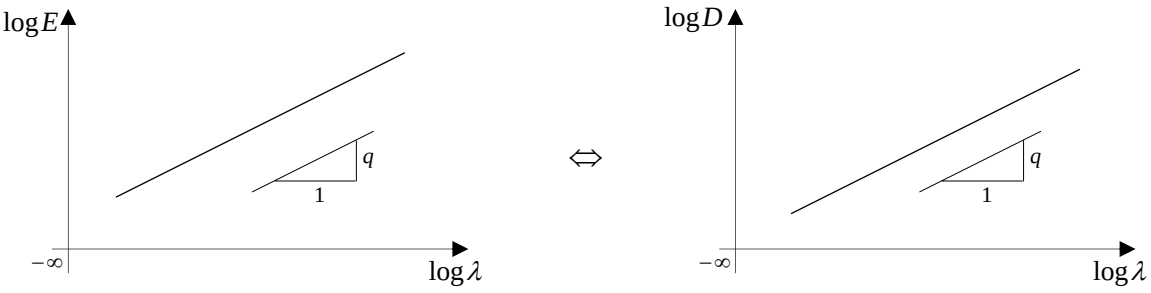

Figure 1. An existing equivalence between errors and pseudo-errors [4, 5]

Eq. (1)), and, considering **U** as the exact solution, the error, *E*, is defined as [8]:

$$E = \left\| \mathbf{U}^a - \mathbf{U} \right\| \tag{2}$$

In view of the academic need in an in-progress research, the main objective in this paper is to display that for two approximate computations for one problem converging properly [7] with identical rates, the computation with more/less error is associated with more/less pseudo-error, and vice versa; see Fig. 2; extension to several computations and even beyond the proper convergence region is considered as the conclusion.

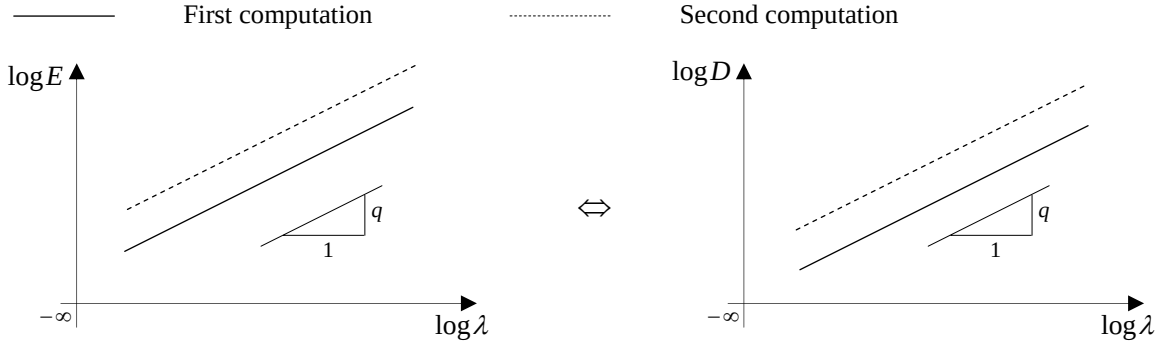

Figure 2. A new equivalence between convergence and pseudo-convergence

Theory

In order to explain the validity of the equivalence addressed in Fig. 2, consider three arbitrary different points on each of the two convergence trends in the left plot in Fig. 2, and address them such that

$$\lambda_1 > \lambda_2 > \lambda_3 > 0 \tag{3}$$

By addressing the two computations with U and $\overline{U}$, and addressing the information about the three points, with the right subscripts "1", "2", and "3, in view of the proper convergence assumption and from simple geometry,

$$\begin{aligned} &\text{For the first computation :} \quad \frac{\log E_1 - \log E_2}{\log \lambda_1 - \log \lambda_2} = \frac{\log E_2 - \log E_3}{\log \lambda_2 - \log \lambda_3} = q \\ &\text{For the second computation :} \quad \frac{\log \overline{E}_1 - \log \overline{E}_2}{\log \lambda_1 - \log \lambda_2} = \frac{\log \overline{E}_2 - \log \overline{E}_3}{\log \lambda_2 - \log \lambda_3} = q \end{aligned} \tag{4}$$

from which,

$$\frac{\log E_1 - \log E_2}{\log \overline{E}_1 - \log \overline{E}_2} = \frac{\log E_2 - \log E_3}{\log \overline{E}_2 - \log \overline{E}_3} = 1 \tag{5}$$

and hence,

$$\begin{aligned} \frac{E_1}{E_2} &= \frac{\overline{E}_1}{\overline{E}_2} \\ \frac{E_2}{E_3} &= \frac{\overline{E}_2}{\overline{E}_3} \end{aligned} \tag{6}$$

Consequently,

$$\frac{E_1}{\overline{E}_1} = \frac{E_2}{\overline{E}_2} = \frac{E_3}{\overline{E}_3} \tag{7}$$

which can be simply extended to (with more points in the left plot in Fig. 2):

$$\frac{E_1}{\overline{E}_1}=\frac{E_2}{\overline{E}_2}=\cdots=\frac{E_n}{\overline{E}_n}=\cdots=\alpha \quad , \quad n\in Z^+ ,\ \alpha\in R^+ \tag{8}$$

By considering different components of the solution separately, and taking into account that, for each component, the proper convergence occurs from one side [9], i.e. with values always more or always less than the exact value, Eq. (1) leads to

$$\begin{aligned} D_i &= \frac{\left|U^{a}{}_i - U^{a}{}_{i-1}\right|}{\left(\frac{\lambda_{i-1}}{\lambda_i}\right)^q - 1} \\ &= \frac{\left|\left(U^{a}{}_i - U\right) - \left(U^{a}{}_{i-1} - U\right)\right|}{\left(\frac{\lambda_{i-1}}{\lambda_i}\right)^q - 1} = \frac{\left|\pm E_i \mp E_{i-1}\right|}{\left(\frac{\lambda_{i-1}}{\lambda_i}\right)^q - 1} = \frac{\left|E_i - E_{i-1}\right|}{\left(\frac{\lambda_{i-1}}{\lambda_i}\right)^q - 1} \end{aligned} \tag{9}$$

for arbitrary component of the solution. From Eqs. (8) and (9), we can conclude that

$$\begin{aligned} D_i &= \frac{\left|E_i - E_{i-1}\right|}{\left(\frac{\lambda_{i-1}}{\lambda_i}\right)^q - 1} \\ &= \frac{\left|\alpha\overline{E}_i - \alpha\overline{E}_{i-1}\right|}{\left(\frac{\lambda_{i-1}}{\lambda_i}\right)^q - 1} = \alpha\frac{\left|\overline{E}_i - \overline{E}_{i-1}\right|}{\left(\frac{\lambda_{i-1}}{\lambda_i}\right)^q - 1} \end{aligned} \tag{10}$$

and since, according to the definition of pseudo-errors (see Eq. (1)), and with attention to [9],

$$\overline{D}_i = \frac{\left|\overline{U}^{a}{}_i - \overline{U}^{a}{}_{i-1}\right|}{\left(\frac{\lambda_{i-1}}{\lambda_i}\right)^q - 1} = \frac{\left|\left(\overline{U}^{a}{}_i - U\right) - \left(\overline{U}^{a}{}_{i-1} - U\right)\right|}{\left(\frac{\lambda_{i-1}}{\lambda_i}\right)^q - 1} = \frac{\left|\pm \overline{E}_i \mp \overline{E}_{i-1}\right|}{\left(\frac{\lambda_{i-1}}{\lambda_i}\right)^q - 1} = \frac{\left|\overline{E}_i - \overline{E}_{i-1}\right|}{\left(\frac{\lambda_{i-1}}{\lambda_i}\right)^q - 1} \tag{11}$$

from Eqs. (8), (10) and (11), we can conclude

$$D_i = \alpha\overline{D}_i \quad \equiv \quad \frac{D_i}{\overline{D}_i} = \alpha , \qquad \alpha\in R^+ \tag{12}$$

Comparing Eqs. (8) and (12) leads to the fact that, when E_i is larger/smaller than $\overline{E}_i$, a similar relation exists between D_i and $\overline{D}_i$, and since this is for an arbitrary component of the solution, it can also be considered valid for the total solutions $\mathbf{U}^a$ and $\overline{\mathbf{U}}^a$ in the regions of proper convergence, i.e. we have succeeded to arrive at the right plot in Fig. 2 from the left plot.

In order to arrive at the left plot in Fig. 2 from the right plot, by considering *n* separate points in the right plot, we can arrive at

$$\frac{D_1}{\overline{D}_1}=\frac{D_2}{\overline{D}_2}=\cdots=\frac{D_n}{\overline{D}_n}=\cdots=\beta \quad , \quad n\in Z^+ ,\ \beta\in R^+ \tag{13}$$

Again restricting the discussion to arbitrary specific component of the solution and considering the fact that properly converging solutions not affected by round off (as displayed in Fig. 2) converge to the exact solution from one side [9], from Eqs. (1) and (13), we can conclude

$$\frac{U^a{}_0 - U^a{}_1}{\overline{U}^a{}_0 - \overline{U}^a{}_1} = \frac{U^a{}_1 - U^a{}_2}{\overline{U}^a{}_1 - \overline{U}^a{}_2} = \frac{U^a{}_2 - U^a{}_3}{\overline{U}^a{}_2 - \overline{U}^a{}_3} = \cdots = \frac{U^a{}_{n-1} - U^a{}_n}{\overline{U}^a{}_{n-1} - \overline{U}^a{}_n} = \cdots = \pm\beta \quad , \quad n \in Z^+ , \ \beta \in R^+ \tag{14}$$

and since theoretically, when disregarding round-off

$$\overline{U}^a{}_\infty = U^a{}_\infty = U \qquad \text{(the exact solution)} \tag{15}$$

from simple mathematics and Eq. (14), we will arrive at

$\forall\ m = 0,1,2,3,\ldots$

$$\frac{U^a{}_m - U}{\overline{U}^a{}_m - U} = \frac{\left(U^a{}_m - U^a{}_{m+1}\right) + \left(U^a{}_{m+1} - U^a{}_{m+2}\right) + \ldots}{\left(\overline{U}^a{}_m - \overline{U}^a{}_{m+1}\right) + \left(\overline{U}^a{}_{m+1} - \overline{U}^a{}_{m+2}\right) + \ldots} = \frac{\sum_{i=0}^{\infty}\left(U^a{}_{m+i} - U^a{}_{m+i+1}\right)}{\sum_{i=0}^{\infty}\left(\overline{U}^a{}_{m+i} - \overline{U}^a{}_{m+i+1}\right)} = \pm\beta \tag{16}$$

leading to

$$\frac{\left|U^a - U\right|}{\left|\overline{U}^a - U\right|} = \beta \tag{17}$$

or

$$\frac{E}{\overline{E}} = \beta \tag{18}$$

implying the completion of the proof for an arbitrary component of the solution and accordingly for the total solution in the proper convergence region. Thus, the proof is complete. As a direct consequence, when a problem can be solved approximately with several methods, and the computational methods provide similar rates of convergence and converge properly, the method with the i th size of error is also of the i th size of pseudo-error; see Fig. 3 for five methods; specifically the method displaying the least pseudo-error leads to the most accurate solution. Furthermore, from Eqs. (8), (12), (13), and (18), for any two of these computations, in the region of proper convergence,

$$\frac{E}{\overline{E}} = \frac{D}{\overline{D}} = \text{Const.} > 0 \tag{19}$$

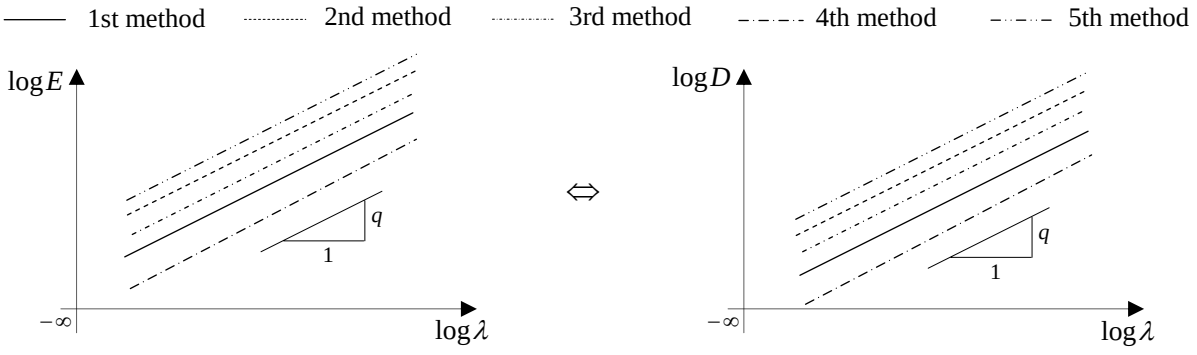

Figure 3. An extension of Fig. 2

And finally, since the basis of convergence plots and specifically the proper convergence regions is the Taylor series expansion [10] of approximate computations with respect to the algorithmic parameters [11], i.e.

$$\mathbf{U}^{a} = \mathbf{U} + \mathbf{C}_0 \lambda^{q} + \mathbf{C}_1 \lambda^{q+1} + \cdots \quad , \quad \mathbf{C}_0 \neq \overline{\mathbf{O}} \tag{20}$$

in view of the continuity of the Taylor series expansion addressed in Eq. (20) with respect to λ [10], it is also reasonable to expect the validity of the claim implied in Eq. (19) and Fig. 3 for values of the algorithmic parameter slightly larger than those corresponding to the proper convergence regions.

An Example

Since the previous section is carried out in a mostly rigorous manner, for the sake of brevity, only one example is presented here (and this example is also studied for other purposes in [12, 13]). Consider the shear frame structural system defined in Fig. 4 and Table 1 ($\ddot{u}_g$ stands for the ground acceleration and $_f\Delta t$ implies the size by which $\ddot{u}_g$ is digitized). Transient analysis of the structural behavior by direct time integration [14-17] is the approximate computation under consideration. The time integration methods are the Houbolt, the average acceleration Newmark, the C-H ($\rho_\infty = 0.8$), and the C-H ($\rho_\infty = 0.5$) methods [18-22], all providing second order convergence, i.e. $q = 2$ [14, 15, 23]. The peak lateral displacement of the top floor and the final floors shear forces are the target solutions. In the study of the convergence trend, the integration steps are the algorithmic parameters [11, 14, 15, 24, 25], and as conventional [17, 26, 27], the steps of direct time integration are halved sequentially,

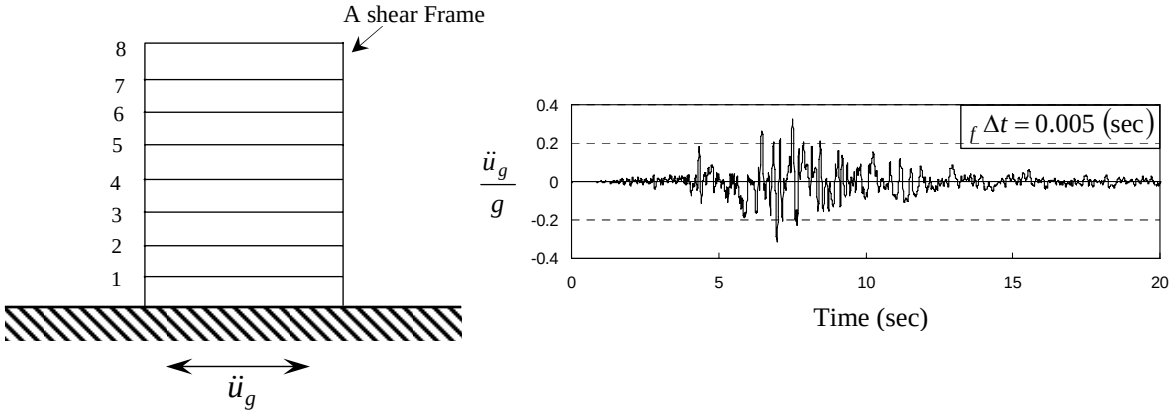

Figure 4: Structural system under consideration: (a) Structural model, (b) Excitation

Table 1: Some properties of the shear frame in Fig. 4(a)

Storey	Mass (Kg)	Stiffness (N/m^2)	Damping (N/m/sec)
1	1036E4	860E7	0
2	1034E4	840E7	
3	1032E4	820E7	
4	1030E4	700E7	
5	1028E4	680E7	
6	1026E4	660E7	
7	1024E4	640E7	
8	1022E4	620E7	

while in the first analysis (computation),

$$\Delta t = \lambda =_f \Delta t = 0.005 \quad \text{(sec)} \tag{21}$$

In a special time integration analysis (computation), carried out with the purpose to determine the errors with high precision, the steps of the direct time integration are considered equal to the very small values stated below:

$$\Delta t = \lambda = 0.005\left(2^{-12}\right) \quad \text{(sec)} \tag{22}$$

The convergence and pseudo-convergence plots are depicted in Figs. 5 and 6, while for the second target solution, i.e. the floors final shear forces, the L_2 norm is implemented for computing the errors and pseudo-errors. Figs. 5 and 6 clearly display the validity of the claims discussed in the previous section, i.e. (1) larger/smaller errors imply larger/smaller pseudo-errors, in the proper convergence regions, and vice versa, (2) validity of Eq. (19) in the proper convergence regions, (3) possibility to extend the previous two points to values of the algorithmic parameter slightly larger than those corresponding to the proper convergence regions.

Several other examples concentrating on different approximate computations, including structural analysis by finite elements, nonlinearity solutions, and different ways of computing π are under study.

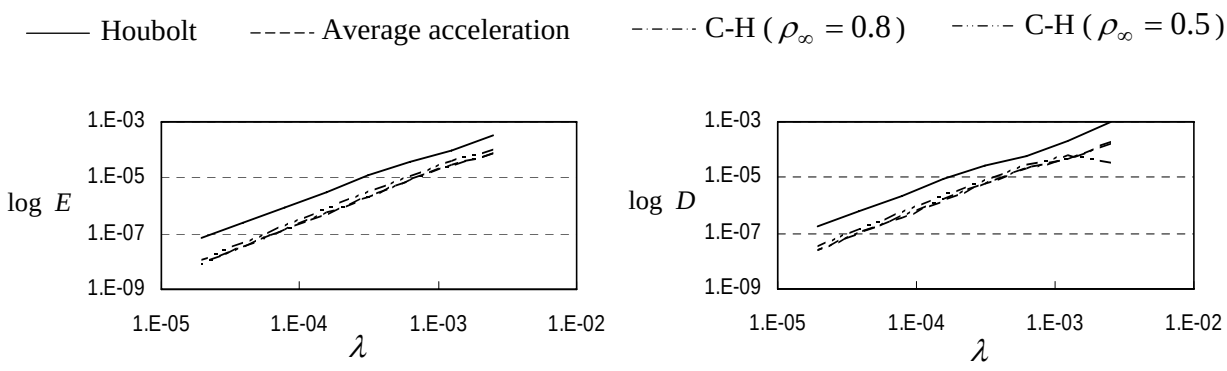

Figure 5. Convergence and pseudo-convergence plots for the peak top displacements

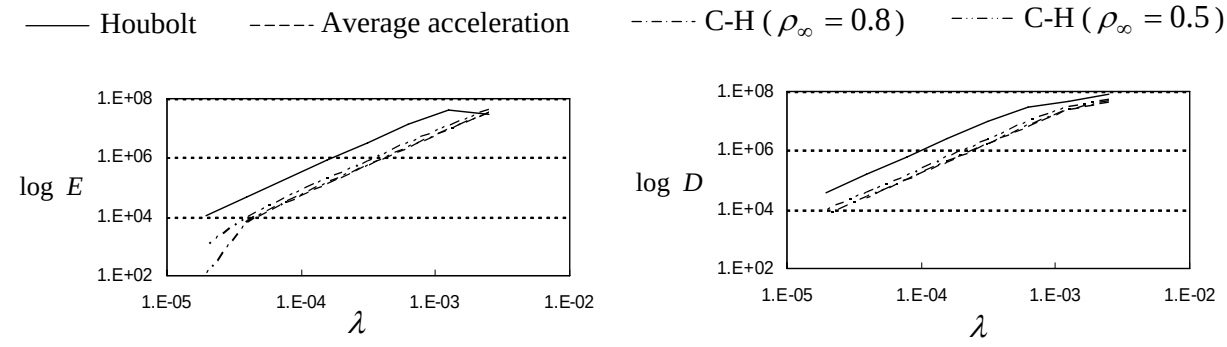

Figure 6. Convergence and pseudo-convergence plots for the final floor shear forces with the L_2 norm

Discussion

Recognition of the most accurate computation from several approximate computations, though of high importance, is generally not an easy task. From the other side of view, arriving at solutions very close to the exact solutions, in order to lead to the computational errors, is computationally expensive. Even the existing error estimations are not reliable in many cases. In these cases, the discussions presented in this paper can be significantly effective.

In this section, some complementary explanation is stated about proper convergence, how to check proper convergence with small computational effort, and meanwhile extension of the discussion to vector and matrix solutions. The equivalence addressed in Fig. 1 is a simple way to check the proper convergence (a simpler way based on purification of convergence [28] is yet not finalized). Nevertheless, for locating each point in the pseudo convergence plot two approximate computations should be carried out and for checking the proper convergence, at least two points should be located. This means three computations, and sounds entailing considerable additional computational cost, especially when taking into account that approximate computations with smaller algorithmic parameters are more costly. This is however not correct. For many approximate computations, e.g. solution of ordinary differential equations, finite element analysis, time integration analysis and nonlinear time history analysis against seismic excitations [17, 26, 27, 29], repetition of the computation after assigning smaller (mostly half) values to the algorithmic parameters (and even repetition of the computation by times) is strongly recommended and in cases considered as an obligatory requirement; see [27]. Therefore, the above-mentioned additional computational effort at most corresponds to one computation, and even the additional cost can be obviated by different approaches, from which, two are (also see [30]): (1) assigning slightly larger values to the algorithmic parameter in the first computation, and (2) while repeating the computation, considering smaller values for r in

$$\frac{\lambda_1}{\lambda_2} = \frac{\lambda_2}{\lambda_3} = \cdots = r \qquad , \qquad r > 1 \tag{23}$$

Consequently, even in cases that additional computational cost basically exists, for practical implementation of the achievements, the additional costs can mostly be hesitated or lessened.

Finally, for non-scalar solutions, the requirement of proper convergence for all components of the solution can be merely sufficient, not necessary (based on the norm). Components of the solution with small contribution in the error/pseudo-error need less to converge properly.

Conclusions

In this paper, it is displayed via theoretical discussion and an example, that for arbitrary approximate computation, solution, and problem:

(1) From several solutions converging properly to the exact solution with similar rates, the most/least accurate solution converges with least/most pseudo-errors, and vice versa.
(2) The ratio of the errors of a computation to the errors of another computation when both computations converge properly and with the same rate remains unchanged, if instead of the errors we compare the pseudo-errors.
(3) The above two points persist for values of the algorithmic parameter slightly larger than those corresponding to the proper convergence regions.

Considering the unavailability of exact solutions and the high computational effort associated with highly precise solutions, implementation of the achievements in practice is briefly discussed; more investigation is essential and strongly recommended.

References

[1] Henrici, P. (1962) *Discrete Variable Methods in Ordinary Differential Equations*, Prentice-Hall, USA.

[2] Srikwerda, J. C. (1989) *Finite Difference Schemes and Partial Differential Equations*, Wadsworth & Books/Cole, USA.

[3] Soroushian, A., *New methods to maintain responses 'convergence and control responses' errors in the analysis of nonlinear dynamic models of structural systems*, PhD Thesis, University of Tehran, Iran, 2003 (in Persian).

[4] Soroushian, A., Pseudo convergence and its implementation in engineering approximate computations, *Proceedings of 4th International Conference from Scientific Computing to Computational Engineering (IC-SCCE 2010)*, Athens, Greece, July 7-10, 2010.

[5] Soroushian, A., Equivalence between convergence and pseudo convergence when algorithmic parameters do not change geometrically, *Proceedings of 6th International Conference from Scientific Computing to Computational Engineering (IC-SCCE 2014)*, Athens, Greece, July 9-12, 2014.

[6] Noble, B., Daniel, J. W. (1977) *Applied Linear Algebra*, Prentice Hall, USA.

[7] Soroushian, A., Proper convergence a concept new in science and important in engineering, *Proceedings of 4th International Conference from Scientific Computing to Computational Engineering (IC-SCCE 2010)*, Athens, Greece, July 7-10, 2010.

[8] Ralston, A., Rabinowitz, P. (1978) *First Course in Numerical Analysis*, McGraw-Hill, USA.

[9] Soroushian, A., Wriggers, P., and Farjoodi, J., A statement for the convergence of approximate responses and its application in structural dynamics, *Proceedings of 2nd International Conference from Scientific Computing to Computational Engineering (IC-SCCE 2006)*, Athens, Greece, July 5-8, 2006.

[10] Apostol, T. M. (1967) *Calculus, Vol. I.*, Wiley, USA.

[11] Soroushian, A., Wriggers, P., Farjoodi, J. (2009) Asymptotic upper bounds for the errors of Richardson extrapolation with practical application in approximate computations, Int J Numer Meth Eng **80**, 565-595.

[12] Soroushian, A., Transient analysis with steps larger than the excitation steps independent from the frequency contents, *Proceedings of 6th ECCOMAS Thematic Conference on Computational Methods in Earthquake Engineering and Structural Dynamics (COMPDYN 2017)*, Island of Rhodes, Greece, June 15-17, 2017.

[13] Soroushian, A., Zarabimanesh, Y., Soleymani, K., Khalkhali, S. M., A new technique for fractional enlargement of integration steps in transient analysis against digitized excitations, *Proceedings of the International Conference on Structural Engineering Dynamics (ICEDyn 2017)*, Ericeira, Portugal, July 3-5, 2017.

[14] Belytschko, T., Hughes, T. J. R. (1983) *Computational Methods for Transient Analysis*, Elsevier, 1983.

[15] Wood, W. L. (1990) *Practical Time Stepping Schemes*, Oxford, 1990.

[16] Paultre, P. (2010) *Dynamics of Structures*, Wiley, USA.

[17] Clough, R. W., Penzien, J. (1993) *Dynamics of Structures*, McGraw-Hill, 1993.

[18] Houbolt, J. C. (1950) A recurrence matrix solution for the dynamic response of elastic aircraft, J Aeronaut Sci **17**, 540–550.

[19] Katona, M. G., Zienkiewicz, O. C. (1985) A unified set of single step algorithms part 3: the beta-m method, a generalization of the Newmark scheme, Int J Numer Meth Eng **21**, 1345–1359.

[20] Soroushian, A., Farjoodi, J. (2008) A unified starting procedure for the Houbolt method, Commun Numer Meth Eng **24**, 1–13.

[21] Newmark, N. M. (1959) A method of computation for structural dynamics, J Eng Mech-ASCE **85**, 67–94.

[22] Chung, J., Hulbert, G. M. (1993) A time integration algorithm for structural dynamics with improved numerical dissipation: the generalized-a method, J Appl Mech-T ASME **60**, 371–375.

[23] Hughes, T. J. R. (1987) *The Finite Element Method: Linear Static and Dynamic Finite Element Analysis*, Prentice-Hall, USA.

[24] Soroushian, A. (2017) Integration step size and its adequate selection in analysis of structural systems against earthquakes. In Eds. M. Papadrakakis, V. Plevris, N.D. Lagaros, eds., *Computational Methods in Earthquake Engineering Vol 3*, Springer, Norway.

[25] Soroushian, A. (2008) A technique for time integration with steps larger than the excitation steps, Commun in Numer Meth Eng **24**, 2087-2111.

[26] Hairer, E., Wanner, G. (1996) *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*, Springer, USA.

[27] Standards New Zealand (2004) *Structural Design Actions Part 5: Earthquake Actions–New Zealand Commentary (Supplement to NZS 1170.5:2004) (2004) NZS 1170.5 Supp 1*, New Zealand.

[28] Soroushian, A., Purification of convergence an approach towards reliable error evaluation, *Proceedings of 11th World Congress on Computational Mechanics (WCCM XI)*, Barcelona, Spain, July 20-25, 2014.

[29] Fish, J., Belytschko, Y. (2009) *A First Course in Finite Elements*, Wiley, UK.

[30] Soroushian, A., A simple approach towards further accuracy in structural dynamic analysis, *Proceedings of 4th ECCOMAS Thematic Conference on Computational Methods in Structural Dynamics and Earthquake Engineering (COMPDYN 2013)*, Kos Island, Greece, June 12-14, 2013.