Error bounds and enclosures: The development of numerical analysis and the impact of the contributions by Lothar Collatz

Götz Alefeld | Michael Plum | Christian Wieners

In 1933 Lothar Collatz published his very first article in *Zeitschrift für Angewandte Mathematik und Mechanik (ZAMM)* on error estimates for finite difference methods for partial differential equations. At that time, numerics meant calculations of approximations by hand and mechanical calculators. Then, in the next decades, parallel to the progress in computer technology, more and more methods were developed. Nevertheless, in all cases the accuracy of numerical approximations is limited, so that at least rough error bounds or, at best, tight enclosures are required for the reliability of the numerical scheme and the validation of the approximate results. Here, we recall the early development of numerical analysis of differential equations, of numerical iterations, and for the approximation of eigenvalues.

**KEYWORDS**
development of numerical analysis, enclosures, error bounds

1 | INTRODUCTION

An analytic solution for differential equations can be found only in special cases, in general only approximations are accessible. Efficient methods to evaluate such approximations were already introduced a long time before computing machines were developed, in particular for applications in mechanics and engineering. From the very beginning 100 years ago, the journal *ZAMM* established as a platform for contributions on models and methods for approximations in mechanics. The contributions were written in German, and overview articles were introduced with abstracts in German, English, French, and Russian.

Parallel to the development of more and more advanced numerical methods, numerical analysis established as an independent area of mathematical research. The main task is the investigation of convergence properties of numerical approximations and to derive bounds for the error which cannot be avoided as long as no analytic solution can be determined in closed form. From the very beginning Lothar Collatz contributed to this research, in particular with many articles in *ZAMM*. His achievements were fundamental at that time, since they build a comprehensive treatment of all aspects in this field:

- Approximation schemes for differential equations
  In the first step, derivatives where approximated numerically, e.g., by finite differences, so that the discretized differential equations can be solved as a finite dimensional system.
- Iterative solution methods
The discretization of partial differential equations yields very large finite dimensional systems which have to be solved only approximately so that the truncation error is smaller than the discretization error, e.g., by suitable iterative schemes.

- Error bounds and enclosures
  In the next step, the accuracy of the approximation is analyzed, e.g., on the basis of Taylor expansions, where the main task is the estimate of the global error depending on the local approximation error of the derivatives in the equation and on a priori estimates for the stability of the discrete solutions.

- Applications
  Stability properties of mechanical structures can be predicted by numerical computations, e.g., by the evaluation of eigenvalues and eigenmodes for different components in the mechanical system. Of course, this requires that the model and the computations are sufficiently accurate.

In the next sections we recall examples for this procedure where together with efficient methods for the computation of a numerical approximation the accuracy of the overall solution process is considered as well. This combination of efficient approximation schemes with a broad technology to control and to estimate the numerical error in order to gain reliability of the overall solution process is still the main challenge in numerical mathematics.

2 | NUMERICAL APPROXIMATION OF SOLUTIONS OF DIFFERENTIAL EQUATIONS

Finite differences on uniform regular grids in 2D were the starting point for approximating elliptic equations in bounded domains. If the boundary is aligned with the grid, the consistency error for standard methods is of second order, but in case of curved boundary, it is reduced if only a simple approximation of the boundary is used. In his first publication in 1933, L. Collatz derived an improved boundary interpolation which yields an approximation which has also been of second order [1], cf. Figure 1.

Meanwhile this is extended to higher order methods, but a simple equidistant interpolation of the boundary gets unstable for high polynomial degrees (the so-called Babuška paradox). In the following, the boundary interpolation has also been transferred to the optimal approximation of general domains with isoparametric finite elements, and more recently with isogeometric analysis (IGA) based on rational B-splines.

Nonlinear ordinary differential equations of first and second order \( \dot{y} = f(t, y) \) and \( \ddot{y} = f(t, y, \dot{y}) \) are considered by Collatz in 1942 [2]. At that time the error analysis for first order ODEs was already developed, e.g., for linear multi-step methods. In [2], the two-step methods by Adams (for the first order case) and Störmer (for second order) are considered.
Today these methods are well understood and many software tools are available, but the main challenge remains to find a compromise between the efficiency of explicit methods and the improved stability of implicit methods: explicit methods are easier to realize but not suitable for stiff problems, since then the stability requires very small time steps. On the other hand, implicit methods require the solution of a nonlinear system of algebraic equations. This can be approximated by a predictor-corrector method, starting with an explicit prediction as a starting point for nonlinear correction steps. Such a method is introduced in [2], the error is estimated and criteria for the optimal step size and the required iterations for every time step are derived. Moreover, for the first time step \( k = 1 \) a starting iteration is defined; this is required, since the two-step method computes \( y_{k+1} \) from \( y_k \) and \( y_{k-1} \).

More involved is the error analysis for the second order case. Here, explicit bounds for the step size depending on the two Lipschitz constants of \( f \) with respect to \( y \) and \( \dot{y} \) are derived. On basis of this analysis, in [3] Collatz introduced a “natural” time step size which is optimal in the sense that the method is stable and convergent and in addition aims at minimizing the computational effort. Here it is recommended to adjust the step size during the simulation in order to remain within a certain range of optimality.

Meanwhile this is a general paradigm for adaptive computations, and a large variety of techniques for error estimation and step size control is available for ODEs as well as for PDEs in time and in space. Nevertheless, uniform step sizes are still dominant in many applications, since step size adaption also requires additional computations and the development of adaptive algorithms is more involved and often problem specific, so that the code has to be extended for new problem classes. Thus, optimality within a prescribed level of accuracy remains a challenge.

In 1949 the method and the error analysis for second order ODEs are extended to higher order ODEs [4]. Here, a general technique is presented for the explicit representation of multiple integrals by finite differences and a remainder term depending of higher derivatives. This allows to construct a class of multi-step schemes for \( n \)-th order ODEs. As an example, one obtains for the order \( n = 4 \), time step size \( h > 0 \) and \( k = 1, 2, 3, 4, \ldots \) the implicit system

\[
\begin{align*}
y^{(3)}_{k+1} &= y^{(3)}_{k-1} + h \left( f_k + \frac{1}{3} \nabla^2 f_k \right), \\
y^{(2)}_{k+1} &= 2y^{(2)}_k - y^{(2)}_{k-1} + h^2 \left( f_k + \frac{1}{12} \nabla^2 f_k \right), \\
y^{(1)}_{k+1} &= y^{(1)}_{k-1} + 2h y^{(2)}_k + h^3 \left( \frac{1}{3} f_k + \frac{1}{60} \nabla^2 f_k \right), \\
y^{(1)}_{k+1} &= 2y_{k} - y_{k-1} + h^2 y^{(2)}_k + h^4 \left( \frac{1}{12} f_k + \frac{1}{360} \nabla^2 f_k \right),
\end{align*}
\]

where \( t_k = kh, f_k = f(t_k, y_k), \) and \( \nabla^2 f_k = f_{k+1} - 2f_k + f_{k-1} \). For given initial values \( y_0, y^{(1)}_0, y^{(2)}_0, y^{(3)}_0 \) a starting iteration is used to compute the values for the first time step \( y_1, y^{(1)}_1, y^{(2)}_1, y^{(3)}_1 \).

Later, a major focus in the research of Collatz is the numerics for partial differential equations. An example [5] in 1952 is the error analysis for approximations of second order elliptic equations in a bounded domain \( \Omega \) with Dirichlet boundary data on \( \partial \Omega \). Here, a polynomial ansatz in \( \Omega \) is used, and the approximation is computed by minimizing a functional combining the corresponding energy and a penalty term for the boundary values. The main tool in this analysis is the monotonicity principle. An example is provided for the Laplace equation \( \Delta u = 0 \) in the unit square for given polynomial Dirichlet data \( u_\Omega \). The best approximation with a polynomial ansatz function

\[
p(x, y) = a_1 + a_2 x^2 + a_3 y^2 + a_4 (x^4 + y^4) + a_5 x^2 y^2
\]

is computed by minimizing the corresponding quadratic functional (with penalty parameter \( \vartheta > 0 \) for the boundary term)

\[
J(p) = \frac{1}{2} \int_\Omega |\nabla p|^2 \, dx + \vartheta \int_{\partial \Omega} |p - u_\Omega|^2 \, da
\]

which is achieved by solving a linear system in \( \mathbb{R}^5 \). Then, an explicit bound for the pointwise error \( \| u - p \|_{\infty, \Omega} \) is derived from evaluating \( \| \Delta p \|_{\infty, \Omega} \) and \( \| p - u_\Omega \|_{\infty, \partial \Omega} \); see also Section 4.
Meanwhile, for the finite element method with piecewise polynomial ansatz functions, minimizing the energy functional is the standard approach, where Dirichlet boundary conditions in general are approximated by interpolation. Nevertheless, in particular for nonconforming approximations such as discontinuous Galerkin methods, the Dirichlet data are approximated using a penalty term. On the other hand, since for many elliptic applications such as linear elasticity no monotonicity principle exists, in general error estimates for finite elements are now derived by Galerkin orthogonality.

The state of the art in numerical analysis of differential equations is compiled by Lothar Collatz in a textbook [6] with the first edition 1951 written in German; in the following several further editions appeared, including an English translation of the last edition [7] in 2012. Here, in four chapters initial-value problems and boundary-value problems for ordinary differential equations, and time-dependent and stationary partial differential equations are considered. This book was a milestone in this field since it built the basis for the first generation of students in mathematics who were educated in numerical analysis at a time where computers were available only in advanced computing centers. Meanwhile both, numerical techniques and computer facilities advanced enormously, but the main message is still the same: numerical mathematics aims simultaneously at both, the development of efficient methods together with the analysis of convergence and error estimates.

3 CONVERGENCE ESTIMATES FOR NUMERICAL ITERATIONS IN \( \mathbb{R} \) AND IN \( \mathbb{R}^N \)

Today, it is generally accepted that for a new numerical method (or for a modification of a well-known method) among others, several aspects have to be discussed:

(a) the numerical behavior with respect to rounding errors (e.g., in a given floating point system);
(b) the necessary amount of work (e.g., the number of floating point operations);
(c) the error caused by stopping an iterative method after a finite number of steps.

Before numerical analysis established as research topic, numerical algorithms were often considered without taking into account (a)–(c). In the development of corresponding ideas and results, the work of Lothar Collatz has played a major role. Here, we demonstrate this by referring to several examples in a series of his articles published in the early time of ZAMM.

Concerning (a), we start with the paper [8] which appeared in 1936, where the computation of an approximation for the square root of a positive number \( a > 1 \) using a mechanical calculator has been discussed. The proposed algorithm delivers automatically an error bound for the approximation. Furthermore, the new algorithm is faster than the well known Heron method which is based on Newton’s method.

The second aspect (b), namely the discussion of the amount of work and the possibility of its reduction is demonstrated in 1940 in the paper [9] on the evaluation of complex polynomials. It is well known that an effective method is the Horner scheme. In some cases one has to evaluate polynomials with real coefficients at complex numbers; as an example, the application to special mechanical systems with several degrees of freedom is mentioned. For this case, an algorithm is presented in detail, which needs half of the arithmetic work compared to applying the Horner scheme directly. No complex arithmetic is used in the new algorithm.

The third aspect (c) is first discussed 1942 in [10] for the so-called Total Step Method (TSM) and the Single Step Method (SSM), respectively. Here the question is investigated how one can modify a given linear system of simultaneous equations to make it sufficiently strictly diagonally dominant for the iterative solution process. For a system with two equations and two unknowns it is demonstrated that TSM and SSM are either both convergent for arbitrary starting values or both divergent. In the first case, SSM is twice as fast convergent as TSM. Furthermore it is demonstrated by examples that for systems with more than two unknowns and equations TSM can be convergent whereas SSM is divergent and vice versa. The matrices are not strictly diagonally dominant in these examples. Finally, for a strictly diagonally dominant matrix an error estimation for SSM is derived for the approximation obtained after a finite number of iteration steps.

In 1950, in the paper [11] the fact is used that an iteration method for the linear system \( Ax = b \) resulting from the splitting \( A = B + C \) with a nonsingular matrix \( B \) and iterating by

\[
Bx^{k+1} = -Cx^k + b, \quad k = 0, 1, 2, \ldots
\]
A reducible matrix

\[
\begin{array}{cc}
A_{11} & A_{12} \\
0 & A_{22}
\end{array}
\]

**FIGURE 2** A reducible matrix

is convergent for all starting values if and only if the spectral radius of the so-called iteration matrix \(-B^{-1}C\) is less than one. It is shown in this paper that TSM and SSM can be written in this form by choosing \(B\) and \(C\) correspondingly. Furthermore, the spectral radius condition holds if \(A\) is strictly diagonally dominant. The same holds if the transposed of \(A\) is strictly diagonally dominant. Additionally, the spectral radius condition holds for SSM if \(A\) is symmetric and positive definite. An important result of this paper concerns the situation that the given matrix (or its transposed) is only weakly diagonally dominant, i.e.,

\[
\sum_{j=1, j\neq i}^{n} |a_{ij}| \leq |a_{ii}|, \quad i = 1, ..., n,
\]

and where at least for one index \(i\) the strict inequality holds. An example is presented which shows that in this case possibly neither TSM nor SSM is convergent for all initial values. An additional condition for the matrix \(A\) is needed to guarantee convergence: The matrix has to be irreducible ("nichtzerfallend"). This means that it is not possible by interchanging the rows of the matrix and simultaneously interchanging the same columns so that the resulting matrix has the form where \(A_{11}\) and \(A_{22}\) are square block matrices (possibly of different size) and the block \(A_{21}\) has only zero entries. Otherwise, the matrix is reducible ("zerfallend"), cf. Figure 2. This basic property played a major role in the development of more general iterative methods for linear systems. A weakly diagonally dominant matrix which is irreducible is called irreducibly diagonally dominant.

Matrices of this type appear by discretizing certain boundary value problems for partial differential equations (see also Section 2), and now these results are addressed in introductory lectures on numerical analysis. Nevertheless, meanwhile Krylov methods such as conjugate gradients or GMRES are mostly used for the iterative solution of linear systems, and more general convergence results are obtained from spectral properties of the iteration matrix.

For solving simultaneous systems of nonlinear equations the application of an iterative method is usually the only possible procedure for the approximation of a solution. In 1954, the paper [12] considers the so-called simplified Newton method. In this case the derivative is kept fixed after a couple of iterative steps, usually after the first step. For this special situation an error estimation is given for the approximation obtained after a finite number of steps. This is based on a systematic use of Banach’s fixed point theorem. Meanwhile, this idea is adapted for proving the existence of solutions and computing error estimates for nonlinear equations formulated in very general settings; see also Section 4.

We note that today, besides of (a)–(c), on modern computer systems further aspects have to be considered for new algorithms or modifications of existing methods (e.g., parallelization, energy saving by minimizing data transfer in the computer and others). It is clear that at the time when ZAMM was founded these problems did not exist and therefore could not even be formulated. However, it is obvious that without the early results challenging questions for modern applications could not be discussed today.

### 4 | SOLUTION ENCLOSURES AND EIGENVALUE BOUNDS

Collatz’ work is under several aspects guided by the mutual influence of analysis and numerics on each other. He used, e.g., operator theoretical approaches to develop numerical methods for differential equations. On the other hand, he exploited numerical results to obtain analytical statements via enclosures of exact solutions, e.g., for boundary and eigenvalue problems with ordinary or partial differential equations. Hence he was a pioneer in the field of “computer-assisted proofs”, which nowadays (with much more powerful computer facilities) is a worldwide rapidly growing area.
We select some early ZAMM publications of Collatz to demonstrate this interplay of analysis and numerics in his work. In [13], he considers fixed-point equations

\[ Tu = u \]  

(1)
in a Banach space \( B \), and also equations \( Su = 0 \), which are transformed into the fixed-point form (1) using \( T = I - L^{-1}S \), where \( L \) denotes an invertible linear operator. Under suitable assumptions one may hope for contraction properties of \( T \) near a desired exact solution, and hence for convergence of the sequence \( (u_n) \) given by some appropriate \( u_0 \) and

\[ u_{n+1} = Tu_n, \quad n = 0, 1, 2, \ldots \]  

(2)

which Collatz uses for obtaining numerical approximations. He also formulates a general contraction mapping theorem for problem (1) which uses a metric with values in an ordered Banach space. As main application, he considers nonlinear differential equations

\[ Lu(x) = f(x, u(x)) + \text{boundary conditions} \]

on a bounded domain \( D \subset \mathbb{R}^m \), where \( L \) is a linear differential operator with suitable bijectivity properties. With \( N \) denoting an upper bound for \( \partial f / \partial u \) on some appropriate set, the crucial assumption that some positive function \( \sigma \) exists such that

\[ L\sigma - N\sigma \geq N|u_1 - u_0| \quad \text{on } D \]  

(3)

provides the rigorous error bound

\[ |u - u_1| \leq \sigma \quad \text{on } D \]  

(4)

for an exact solution \( u \).

Assumption (3) is closely related to the existence of a maximum principle for \( L - N \) and hence rather restrictive, but nevertheless Collatz and others were able to obtain solution enclosures by this or similar approaches for many examples with ordinary or partial differential equations. Meanwhile, since about 30 years, methods are being developed which replace (4) by an error bound in a Sobolev space and which avoid condition (3).

Collatz’ approach via (3) is also strongly connected with methods of monotone operators, which he developed and applied to many examples in various publications, e.g., [13,14]. Here, the Banach space \( B \) in which the fixed-point Equation (1) is formulated is assumed to be ordered, and \( T \) is supposed to be either monotonically non-decreasing or monotonically non-increasing with respect to this order. In the first case, besides (2) a second sequence \( (\hat{u}_n) \), \( \hat{u}_{n+1} = T\hat{u}_n \), is considered, and under the assumptions

\[ u_0 \leq \hat{u}_0, \quad u_0 \leq u_1, \quad \hat{u}_1 \leq \hat{u}_0, \]  

(5)

the existence of a fixed-point \( u^* \) follows from Schauder’s fixed-point theorem; both sequences converge to \( u^* \), and

\[ u_0 \leq u_1 \leq \cdots \leq u_n \leq u^* \leq \hat{u}_n \leq \cdots \leq \hat{u}_1 \leq \hat{u}_0 \quad \text{for all } n \in \mathbb{N}. \]

Condition (5) is also closely related to the method of sub- and supersolutions for enclosing (and proving existence of) solutions to initial and boundary value problems.

In the second case, where \( T \) is monotonically non-increasing, the assumptions \( u_0 \leq u_1 \) and \( u_0 \leq u_2 \) lead to the convergence of \( (u_{2k}) \) and \( (u_{2k+1}) \) to limits \( u_{\text{even}} \) and \( u_{\text{odd}} \), respectively, and the existence of a fixed-point \( u^* \) such that

\[ u_0 \leq u_2 \leq u_4 \leq \cdots \leq u_{\text{even}} \leq u^* \leq u_{\text{odd}} \leq \cdots \leq u_5 \leq u_3 \leq u_1. \]
Another major topic in Collatz’ work on solution enclosures is the field of eigenvalue bounds. E.g., in [15], he considers eigenvalue problems

\[ L[\varphi] + \lambda p \varphi = 0 \quad \text{boundary conditions} \quad (6) \]

with a symmetric, negative definite linear differential operator \( L \) and a positive, continuous function \( p \). In some parts of [15], \( p \varphi \) in (6) is replaced by a second differential operator \( M[\varphi] \). In [16], also integro-differential operators \( L \) and \( M \) are considered. Under suitable assumptions, Collatz uses an inverse iteration \( L[F_n] + p F_{n-1} = 0 \) to compute approximations to the eigenfunction associated with the first (smallest) eigenvalue \( \lambda_1 \), and then evaluates the Rayleigh quotient to obtain approximations to \( \lambda_1 \), which by Rayleigh’s principle also form upper bounds to \( \lambda_1 \). To compute lower bounds to \( \lambda_1 \), Collatz proposes the Temple quotients for \( F_1, F_2, \ldots \), i.e.,

\[
\lambda_1 \geq \frac{\ell_2 \int F_n(-L)[F_n]dx - \int \frac{1}{p} L[F_n]^2 dx}{\ell_2 \int p F_n^2 dx - \int F_n(-L)[F_n]dx} \quad \text{for all } n \in \mathbb{N}, \quad (7)
\]

with \( \ell_2 \) denoting a (rough) lower bound for \( \lambda_2 \), which must be large enough to make the denominator in (7) positive. Actually, the bound (7) is obtained by applying the formula [15, Equation (7.4)] for odd \( n \).

For computing upper bounds also for higher eigenvalues, Collatz proposes the Rayleigh-Ritz method, which he formulates in two versions, once via the Euler equations of a variational problem, and once directly via a Ritz-Galerkin approach.

80 years after Collatz’ early publication [15], the Rayleigh-Ritz method and the Temple quotient are still “modern” and widely used. Some other approaches providing lower eigenvalue bounds (Trefftz method, Temple’s inclusion theorem, difference methods), which are also discussed in Collatz’ work, have got a bit in the background since then.

Further developments were Kato’s eigenvalue bounds (1949) which are highly accurate but need substantial a priori knowledge, Lehmann’s lower eigenvalue bounds (1963) and Goerisch’s extension which generalize the Temple quotient (7) for obtaining lower bounds also to higher eigenvalues, and homotopy methods for computing rough lower bounds to some higher eigenvalues, which are needed for the Temple quotient and for Lehmann’s method.

5 CONCLUSION

This review on early developments in numerical mathematics with a special focus on the contributions of Lothar Collatz in ZAMM highlights the fundamental principles in applied mathematics:

- Mathematical models are investigated which are based on applications in mechanics and physics, and in many cases these applications initiate developments in numerics; examples are the computation of eigenmodes of membranes [17], eigenfrequencies of machine components [18] and of plates [19].
- The development of algorithms is complemented by the analysis of stability, accuracy, efficiency, and – as far as possible – error control.
- The approximation of mathematical models and the analysis of the solution properties with respect to existence and regularity have to be combined to guarantee convergence of the approximations.

REFERENCES


How to cite this article: Alefeld G, Plum M, Wieners C. Error bounds and enclosures: The development of numerical analysis and the impact of the contributions by Lothar Collatz. Z Angew Math Mech. 2020;100:e202002042. https://doi.org/10.1002/zamm.202002042