M. N. Spijker

Optimum error estimates for finite-difference methods

Acta Universitatis Carolinae. Mathematica et Physica, Vol. 15 (1974), No. 1-2, 159--164

Persistent URL: http://dml.cz/dmlcz/142348

Optimum Error Estimates for Finite-Difference Methods

M. N. SPIJKER
Mathematisch Instituut, Rijksuniversiteit, Leiden

Estimates for the accumulated error in finite-difference methods are presented which, in a qualitative sense, cannot be improved. These optimum error estimates can be used to derive two-sided estimates for the accumulated error. Besides they are an appropriate tool in comparing the stability of different methods for solving a given differential equation.

I.· Introduction

Let u denote the solution of a finite-difference equation approximating a given differential equation. If the finite-difference equation is perturbed by a quantity w, e.g. due to round-off error, then instead of u we obtain a solution $\tilde{u}$. It is an important task of numerical analysis to establish upper bounds for $\|\tilde{u} - u\|$ in terms of $w, \|\,.\,\|$ denoting an appropriate seminorm.

In this paper we shall deal with optimum error estimates, i.e. estimates of $\|\tilde{u} - u\|$ which, in a qualitative sense, cannot be improved. The proof of the results stated below can be found in the publications [1], [2]. In the following we shall confine ourselves to finite-difference methods for solving initial value problems for first order ordinary differential equations. For further generalizations and application s we refer to the publications listed at the end of this paper.

2. An Optimum Error Estimate for Euler's Method

2.1. Consider the numerical solution of the initial value problem

$$U'(t) = f(t, U(t))\,(0 \leq t \leq T),\ U(0) = s \tag{2.1}$$

by Euler's method

$$u_0 - s = 0,\ h^{-1}(u_n - u_{n-1}) - f(t_{n-1}, u_{n-1}) = 0\ \ (n = 1, 2, \ldots, N), \tag{2.2}$$

where u_n is an approximation of $U(t)$ at $t = t_n = nh$ and the integer N satisfies $Nh = T$. Let

$$\tilde{u}_0 - s = w_0,\ h^{-1}(\tilde{u}_n - \tilde{u}_{n-1}) - f(t_{n-1}, \tilde{u}_{n-1}) = w_n\ \ (n = 1, 2, \ldots, N), \tag{2.3}$$

where w_n denote arbitrary local perturbations (e.g. caused by round-off in the actual application of Euler's method).

Throughout this article, with the exception of section 3.1, we assume that the real function f occuring in (2.1) has domain $[0, T] \times R$ and satisfies a Lipschitz condition

$$|f(t, \tilde{\xi}) - f(t, \xi)| \leq \lambda \, . \, |\tilde{\xi} - \xi| \tag{2.4}$$

for all $t, \xi, \tilde{\xi}$ with $0 \leq t \leq T$ and $\xi, \tilde{\xi} \in R$ (with R we denote the set of real numbers). Using (2.4) the two error estimates

$$\max_{0 \leq n \leq N} |\tilde{u}_n - u_n| \leq \gamma_1 \, . \max_{0 \leq n \leq N} |w_n| \, , \tag{2.5}$$

$$\max_{0 \leq n \leq N} |\tilde{u}_n - u_n| \leq \gamma_2 \, . \, \{|w_0| + h \sum_{i=1}^{N} |w_i|\} \tag{2.6}$$

can be derived. (2.5) and (2.6) hold for arbitrary $N \geq 1$ and w_n, with constants γ_1, γ_2 independent of N and w_n. Note that, in general, u_n, $\tilde{u}_n$ and w_n not only depend on n but also on N.

2.2. The following definitions enable us to compare the structures of error estimates like (2.5) and (2.6).

Definition 1. Let Φ and Φ' be functionals from $\bigcup_{N=1}^{\infty} R^{N+1}$ into R with the property that there exists a constant $\beta > 0$ such that

$$\Phi[w] \leq \beta \, . \, \Phi'[w] \text{ (for all } N \geq 1 \text{ and}$$
$$\text{all } w = (w_0, w_1, \ldots, w_N) \in R^{N+1}) \, .$$

Then we use the notation

$$\Phi \prec \Phi' \, .$$

In the following we consider a fixed initial value problem (2.1) and we consider variable $N \geq 1$, $w_n \in R$ $(0 \leq n \leq N)$.

Consider two arbitrary error estimates

(A) $$\max_{0 \leq n \leq N} |\tilde{u}_n - u_n| \leq \gamma \, . \, \Phi[w]$$

and

(A′) $$\max_{0 \leq n \leq N} |\tilde{u}_n - u_n| \leq \gamma' \, . \, \Phi'[w]$$

valid for all $N \geq 1$ and all $w = (w_0, w_1, \ldots, w_N)$, γ and γ' denoting constants independent of N and w.

Definition 2. The error estimate (A) is *better* than (A′) if $\Phi \prec \Phi'$ and not $\Phi' \prec \Phi$. The estimate (A) is *optimum* if for any other estimate of type (A′) we have $\Phi \prec \Phi'$.

Example. Defining $\Phi_1[w] = \max_{0 \leq n \leq N} |w_n|$, $\Phi_2[w] = |w_0| + h \sum_{i=1}^{N} |w_i|$ we have $\Phi_2 \prec \Phi_1$ but not $\Phi_1 \prec \Phi_2$.
Thus according to definition 2 the estimate (2.6) is better than (2.5).

We have

Theorem 1. The error estimate

$$\max_{0\leq n\leq N} |\tilde{u}_n - u_n| \leq e^{\lambda T} . \max_{0\leq n\leq N} |w_0 + h \sum_{i=1}^{n} w_i| \tag{2.7}$$

is optimum.

(We use the convention $\sum_{i=a}^{b} \dots = 0$ if $a > b$.)

2.3. We briefly mention several consequences and applications of the optimum error estimate (2.7). Define the functional Φ_0 in such a way that the righthand member of (2.7) can be written as $e^{\lambda T} . \Phi_0[w]$.

a) From theorem 1 the nontrivial result follows that there exists an optimum error estimate with a functional $\Phi = \Phi_0$ that is *independent of the differential equation under consideration.*

b) Since (A') evidently holds with $\gamma' = 1$, $\Phi'[w] = \max_{0\leq n\leq N} |\tilde{u}_n - u_n|$ and (2.7) is optimum, we have $\Phi_0 \prec \Phi'$. Hence there is a constant $\beta > 0$ such that

$$\frac{1}{\beta} \max_{0\leq n\leq N} \left| w_0 + h \sum_{i=1}^{n} w_i \right| \leq \max_{0\leq n\leq N} |\tilde{u}_n - u_n| . \tag{2.8}$$

It turns out that in (2.8) β can be taken equal to $\beta = 1 + \lambda T$. Thus (2.7), (2.8) provide us with a *two-sided error estimate* for Euler's method.

c) Suppose $\tilde{u}_n = U(t_n)$. Then (2.7), (2.8) yield a two-sided estimate of the so-called *global discretization error.*

d) Suppose $\tilde{u}_n =$ the approximation obtained by actual calculation on a computer. Then w_n are so-called local round-off errors and (2.7), (2.8) yield a two-sided estimate of the *accumulated round-off error.*

3. Generalizations

3.1. The definitions of chapter 2 are easily extended to the case where Euler's method is used for solving a system of say M first order ordinary differential equations, the main alteration being that $|\xi|$ now stands for a norm of $\xi \in R^M$ instead of the absolute value of the real number ξ. With this modification theorem 1 still holds.

3.2. Definition 2 of chapter 2 can be generalized by replacing

$\max_{0\leq n\leq N} |\tilde{u}_n - u_n|$ in (A) and (A') by

$\max_{0\leq n\leq N} \delta_{N,n} |\tilde{u}_n - u_n|$ where $\delta_{N,n}$ are arbitrary weights ≥ 0.

Let
$$\delta = \sup h \sum_{i=0}^{n-1} \delta_{N,n}/\delta_{N,i}$$
where $h = T/N$ and the supremum is for $N = 1, 2, 3, \ldots$ and $n = 1, 2, \ldots, N$. (In case one of the $\delta_{N,i}$ vanishes we use the conventions $a/0 = \infty$ for $a > 0$, $a/0 = 0$ for $a = 0$.) With the generalized notion of optimum error estimate obtained by the introduction of the weights $\delta_{N,n}$ we have

Theorem 2. Let arbitrary weights $\delta_{N,n} \geq 0$ be given. Then the two propositions (i) and (ii) are equivalent:

(i) $\delta < \infty$;

(ii) There exists a functional Φ with the following property: For any given initial value problem (2.1) there is a constant γ (which may depend on the given initial value problem but not on $N \geq 1$ or $w = (w_0, w_1, \ldots, w_N)$) such that $\max\limits_{0 \leq n \leq N} \delta_{N,n}|u_n - \tilde{u}_n| \leq \gamma \, . \, \Phi[w]$ is an optimum error estimate for Euler's method (applied to the given initial value problem).

This theorem thus expresses the remarkable fact that for any given initial value problem (2.1) there exists an optimum error estimate *with a functional Φ independent of the given initial value problem* if and only if the weights satisfy condition (i).

As an illustration of the above we consider the weights $\delta_{N,n} = (n + 1)^p$ where $-\infty < p < \infty$. A little calculation shows that for $p < 1$ we have $\delta < \infty$, while for $p \geq 1$ we have $\delta = \infty$.

It turns out that, for $p < 1$, the optimum error estimate whose existence is guaranteed by theorem 2, is of the form
$$\max_{0 \leq n \leq N} (n + 1)^p|\tilde{u}_n - u_n| \leq \gamma . \max_{0 \leq n \leq N} (n + 1)^p \, . \, |w_0 + h \sum_{i=1}^{n} w_i| \, .$$
For $p = 0$ this error bound could have been obtained from theorem 1. On the other hand, for $0 < p < 1$, we may deduce from this error bound that
$$|\tilde{u}_N - u_N| \leq \gamma . \max_{0 \leq n \leq N} \left(\frac{n+1}{N+1}\right)^p . \, |w_0 + h \sum_{i=1}^{n} w_i| \, .$$
Note that this estimate of $|\tilde{u}_N - u_N|$ is better than the analogous one obtainable from theorem 1.

3.3. The notions of chapter 2 are easily extended to deal with more general methods for solving (2.1) (e.g. Runge-Kutta methods or linear multistep methods). To this end it is sufficient to replace in chapter 2 the relations (2.2) and (2.3) by (3.1), (3.2), respectively:
$$u_j - s_j = 0, \;\; h^{-1} \sum_{i=0}^{k} \alpha_i u_{n-k+i} - F_n(u_{n-k}, \ldots, u_{n-1}, u_n; \, h) = 0 \, , \tag{3.1}$$
$$\tilde{u}_j - s_j = w_j, \;\; h^{-1} \sum_{i=0}^{k} \alpha_i \tilde{u}_{n-k+i} - F_n(\tilde{u}_{n-k}, \ldots, \tilde{u}_{n-1}, \tilde{u}_n; \, h) = w_n \tag{3.2}$$
where $j = 0, 1, \ldots, k - 1$ and $n = k, k + 1, \ldots, N$.

In (3.1), (3.2) k is a fixed integer ≥ 1, the α_i are real constants (independent of f) with

$$\alpha_k = 1\,, \qquad \alpha_0 + \alpha_1 + \ldots + \alpha_k = 0$$

and $F_n(x_0, x_1, \ldots, x_k;\, h)$ denotes a real function (depending on f) satisfying a uniform Lipschitz condition with respect to its first $k+1$ variables $x_0, x_1, \ldots, x_k$. The s_j are starting values obtained, for instance, by a Taylor expansion of $U(t)$ at $t = 0$.

It is easily verified that Runge-Kutta methods and linear multistep methods as well as many other well known methods can be written in the form (3.1). The generalized notion of an optimum error estimate thus obtained for these methods will be fundamental in the next chapter.

4. Comparing the Stability of Different Methods for Solving the Initial Value Problem (2.1)

4.1. In most current definitions of the concept of stability for finite-difference methods it is required that the error $\tilde{u} - u$ resulting from a local perturbation w admits a bound of the form

$$\max_{0 \leq n \leq N} |\tilde{u}_n - u_n| \leq \gamma \,.\, \Phi[w]\,. \tag{4.1}$$

Depending on the structure of the functional Φ we thus have different concepts of stability.

Let (2.1) be a given initial value problem. Let M_1 and M_2 denote two different methods of type (3.1) for solving it. Then it is natural to call method M_1 more stable than M_2 if M_1 fulfils a stability requirement of type (4.1) which is stronger than any stability requirement fulfilled by method M_2. Since a stronger stability requirement corresponds to a "smaller"functional Φ we are led to the following definition.

Definition 3. Method M_1 is *more stable* than M_2 if there exists an optimum error estimate

$$\max_{0 \leq n \leq N} |\tilde{u}_n - u_n| \leq \delta_1 \,.\, \psi_1[w]$$

for M_1 and an optimum error estimate

$$\max_{0 \leq n \leq N} |\tilde{u}_n - u_n| \leq \delta_2 \,.\, \psi_2[w]$$

for M_2 such that

$$\psi_1 \prec \psi_2 \quad \text{but not} \quad \psi_2 \prec \psi_1\,.$$

4.2. The following definition will be useful to formulate a condition under which it can easily be decided whether a method M_1 is more stable than a method M_2.

Definition 4. A method of type (3.1) is said to satisfy the *strong root condition* if $\xi = 1$ is a simple root of the equation $\alpha_0 + \alpha_1\xi + \ldots + \alpha_k\xi^k = 0$ and all other complex roots ξ have a modulus $|\xi| < 1$.

Theorem 3. Let M_1 and M_2 be two methods of type (3.1) for solving the initial value problem (2.1). Assume M_1 satisfies the strong root condition. Then M_1 is more stable than M_2 if and only if M_2 violates the strong root condition.

Example. Let M_1 stand for the Adams-Bashforth method

$$u_j - s_j = 0,\ h^{-1}(u_n - u_{n-1}) - [3f(t_{n-1}, u_{n-1}) - f(t_{n-2}, u_{n-2})]/2 = 0$$

and M_2 for the midpoint rule

$$u_j - s_j = 0,\ h^{-1}(u_n - u_{n-2}) - 2f(t_{n-1}, u_{n-1}) = 0$$

($j = 0$, 1 and $n = 2, 3, \ldots, N$).

Both methods are of type (3.1) with $k = 2$. Since M_1 satisfies the strong root condition and M_2 does not, the Adams-Bashforth method thus turns out to be more stable than the mid-point rule.

References

[1] SPIJKER, M. N.: On the Structure or Error Estimates for Finite-difference Methods. Numer. Math. 18, 73—100 (1971).

[2] SPIJKER, M. N.: The Existence of Optimum Error Estimates in the Numerical Solution of Differential Equations. (In preparation.)

[3] STETTER, H. J.: Analysis of Discretization Methods for Ordinary Differential Equations. Springer-Verlag, Berlin, Heidelberg, New York (1973).

[4] STUMMEL, F.: Approximation Methods in Analysis. Matematisk Institut, Aarhus Universitet (1973).