Hans J. Stetter

Algorithms for the inclusion of solutions of ordinary initial value problems

In: Jaromír Vosmanský and Miloš Zlámal (eds.): Equadiff 6, Proceedings of the International Conference on Differential Equations and Their Applications held in Brno, Czechoslovakia, Aug. 26 - 30, 1985. J. E. Purkyně University, Department of Mathematics, Brno, 1986. pp. 85--94.

Persistent URL: http://dml.cz/dmlcz/700166

ALGORITHMS FOR THE INCLUSION OF SOLUTIONS OF ORDINARY INITIAL VALUE PROBLEMS

H. J. STETTER
Technical University Vienna
A-1040 Wien, Austria

Introduction

Customary numerical algorithms do not produce bounds for the true solution of the specified problem but an approximate solution. Information about the remaining error is obtained from a *secondary problem*:

Given the original problem and an approximate solution, find an approximation to its error.

It is obvious that this does not eliminate the uncertainty about the quality of the approximate solution. This is tolerable because most problems are only approximations of real-life situations. Nevertheless, there arise situations where rather concise information about the error of an approximate solution must be obtained. In the following, we will analyze the structure of this task for initial value problems for systems of first order ordinary differential equations.

The Problem

We formulate our task in analogy to the secondary problem above. The original problem is $(y(t) \in \mathbb{R}^s)$

$$y'(t) = f(t,y(t)), \quad y(0) = y_0 , \quad t \in [0,T] , \tag{1.1}$$

with sufficient regularity in a sufficiently large neighborhood of the unique solution trajectory $(t,y(t))$.

Inclusion problem: Given an approximate solution $\tilde{y} : [0,T] \to \mathbb{R}^s$ of (1.1)

Find $\tilde{E} : [0,T] \to \mathbb{P}\mathbb{R}^s$ such that

$$e(t) := \tilde{y}(t) - y(t) \in \tilde{E}(t) , \quad t \in [0,T] . \tag{1.2}$$

Here $\mathbb{P}$ denotes the power set; normally we have to restrict the range of

$\tilde{E}$ to an easily representable subset of $\mathbb{IR}^S$ like the set $\mathbb{II}\mathbb{R}^S$ of all intervals in $\mathbb{R}^S$. While the computation of norm bounds for e is a special case of (1.2), we will primarily be interested in componentwise lower and upper bounds which may well be of equal sign. Often we will be satisfied with producing values of $\tilde{E}$ at a sequence of arguments $t_0, t_1, .., t_n \in [0,T]$.

It is clear that the inclusion (1.2) of e implies an inclusion

$$y(t) \in \tilde{y}(t) - \tilde{E}(t) \tag{1.3}$$

for the true solution y of (1.1). The algorithms which we will consider are also immediately applicable to the case of *strips* of true solutions ($y(t) \in \mathbb{IR}^S$) as they appear for a set-valued initial condition $y(0) \in Y_0 \in \mathbb{IR}^S$ in (1.1).

Naturally, the inclusion problem becomes the more delicate the tighter an inclusion we request. It is clear, however, that we cannot generate an inclusion of a prespecified maximal width in a one-pass step-by-step procedure for a general initial value problem.

The generation of numerical solutions for the inclusion problem (1.2) has been studied by many scientists and a good number of algorithms have been proposed. One of the early investigations is by N.J. Lehmann [4]; it is remarkable that he has already suggested the use of symbol manipulation systems in this connection.

For lack of space, we cannot systematically list and comment the various contributions. A very extensive bibliography on the subject is to be found, e.g., in Nickel [7]. Our own bibiliography contains only some typical examples of specific approaches.

Rather than sketching the historic development, we will attempt to display a common conceptual framework for most of the algorithms which have been proposed. This should help in their understanding and evaluation and stimulate the further analysis and development of the area.

Local Analysis

Except in trivial situations, a numerical algorithm for (1.2) cannot cover the interval [0,T] at once. Hence we consider at first *one step* in a forward stepping algorithm: We have arrived at $t_{\nu-1}$ and obtained a set $\tilde{E}_{\nu-1}$ such that $e(t_{\nu-1}) \in \tilde{E}_{\nu-1}$. In the construction of a corresponding

set $\tilde{E}_\nu$ at $t_\nu = t_{\nu-1}+h_\nu$, we have to regard *all* solutions $y(t;t_{\nu-1},y_{\nu-1})$ of (1.1) which pass through an admissible value $y_{\nu-1}$ at $t_{\nu-1}$.

Local problem: Find $\tilde{E}_\nu$ such that (see Fig.1)

$$\tilde{y}(t_\nu) - y(t_\nu;t_{\nu-1},y_{\nu-1}) \in \tilde{E}_\nu \quad \text{for all } y_{\nu-1} \in \tilde{y}(t_{\nu-1}) - \tilde{E}_{\nu-1} \ . \tag{2.1}$$

Obviously, the use of the *local exact inclusion*

$$E_\nu := \{\tilde{y}(t_\nu)-y(t_\nu;t_{\nu-1},y_{\nu-1}) : y_{\nu-1} \in \tilde{y}(t_{\nu-1})-E_{\nu-1}\} \tag{2.2}$$

for $\tilde{E}_\nu$ would keep the inclusion optimally tight. By (2.1), we have $E_\nu \subset \tilde{E}_\nu$ and we can use the *interior difference*

$$D_\nu := \tilde{E}_\nu \dot{-} E_\nu \in I\!P\mathbb{R}^s \tag{2.3}$$

to represent the excess of $\tilde{E}_\nu$ over E_ν.

(For two sets in a linear space, with $A \subset B$, the interior difference $B \dot{-} A$ is the unique set C which satisfies $A + C = B$. Obviously, $0 \in B \dot{-} A$. The norm of $B \dot{-} A$ is the Hausdorff distance of A and B.)

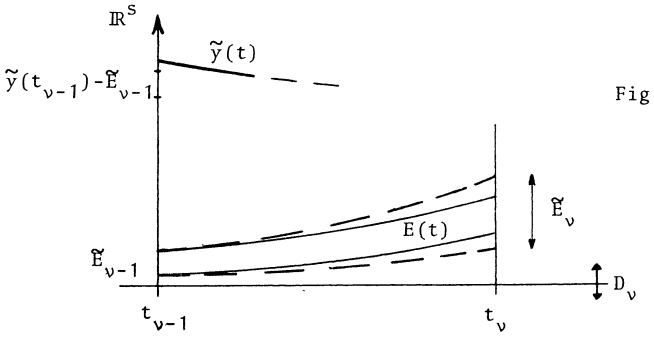

Fig. 1

It appears that the *local excess* D_ν of (2.3) is the natural analogon to the *local error* in a stepwise algorithm for (1.1). Its size, expressed e.g. by

$$\|D_\nu\| := \max_{d \in D_\nu} \|d\|$$

may be used as a quantitative measure of the (local) accuracy of an inclusion algorithm. For $s > 1$ and $\tilde{E}_{\nu-1}, \tilde{E}_\nu \in I\!I\mathbb{R}^s$, D_ν will generally not be an interval because $E_\nu \notin I\!I\mathbb{R}^s$.

Typically, the computation of $\tilde{E}_\nu$ will be based, at best, on

- correct representation of a few derivatives w.r.t. t
- correct representation of *linear* terms in the deviation e between $\tilde{y}$ and y (first order perturbation analysis)
- strict bounding of remainder terms, nonlinearities, etc.

Round-off error effects will be caught by the use of directed rounding. We assume that their influence is negligible compared to the leading local excess terms.

A *local excess analysis* for an algorithm of this kind leads to

$$D_\nu = O(h_\nu^{p+1}) + O(h_\nu \varepsilon_{\nu-1}^2) + \text{higher order terms} \qquad (2.5)$$

where $\varepsilon_{\nu-1} := \operatorname{diam} \tilde{E}_{\nu-1}$. The appearance of the second term seems unavoidable, even if quadratic terms in e are evaluated:

Take $y'=y^2$ so that $y'(y_0+e) = y_0^2 + 2y_0 e + e^2$. Assume $e \in [-\varepsilon,\varepsilon] =: E$ and compute bounds for $y'(y_0+e)$ by interval evaluation of $y_0^2 + 2y_0 E + E^2$. With $E^2 = [0,\varepsilon^2]$, one obtains

$$hy'(y_0+e) \in h \cdot \left[y_0^2 - 2y_0\varepsilon,\ y_0^2 + 2y_0\varepsilon + \varepsilon^2\right]$$

whose lower bound creates an excess of $-h\varepsilon^2$. The reason is the *dependence* between the multiple occurences of E in a quadratic expression.

Stability

If our algorithm accounts for the linear terms (linearized about $\tilde{y}(t)$) in the deviations correctly, the excess D_ν generated in the step towards t_ν should propagate like a local perturbation at t_ν during the further integration. In other words, our local excesses should accumulate like the local errors in a one-step algorithm for (1.1), at least for sufficiently small steps h_ν and a sufficiently narrow inclusion strip.

However, computationally there arises the necessity to *represent* inclusions in terms of a *semiorder* of the $\mathbb{R}^s$ *based on components*, e.g. by componentwise intervals. Not always such a semiorder is preserved by the differential system (1.1): The initial value problem (1.1) is called *quasimonotone* w.r.t. a semiorder $\geqslant$ in $\mathbb{R}^s$ if

$$w'(t) \geq f(t,w(t)) \ , \quad t \in [0,T] \ , \qquad \text{implies} \quad w(t) \geq y(t), \ t \in [0,t] \ .$$
$$w(0) \geq y_0$$

(Criteria for quasimonotony and related theorems may be found in Walter [9].)

If (1.1) is *not* quasimonotone w.r.t. componentwise semiorder, the following happens (see Fig.2):

Consider an inclusion interval $\tilde{E}_{\nu-1} = \tilde{e}_{\nu-1} + [-1,+1] \cdot d_{\nu-1}$, with $d_{\nu-1} \geq 0$. The variational equation of (1.1) (near the solution trajectory) maps $\tilde{E}_{\nu-1}$ into $E_\nu = G\,\tilde{E}_{\nu-1} + \ldots$, but E_ν is not represented or included by $e_\nu + [-1,+1] \cdot G d_{\nu-1}$.

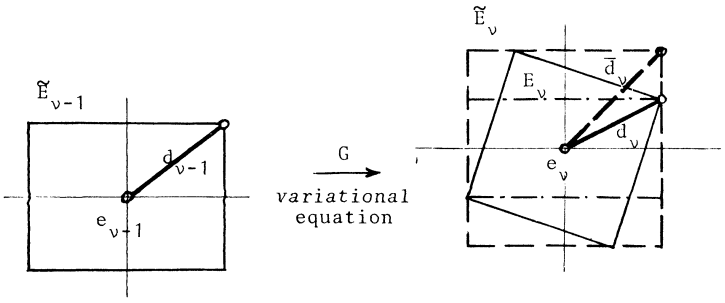

Fig. 2

The smallest interval including E_ν is rather specified by

$$\bar{d}_\nu = |G| \cdot d_{\nu-1} \ . \tag{3.1}$$

Since $G \approx I + h_\nu J_{\nu-1}$, (3.1) simulates the use of the modified Jacobian

$$J^+_{\nu-1} := \begin{pmatrix} \diagdown & & |R| \\ & D & \\ |R| & & \diagdown \end{pmatrix} \tag{3.2}$$

i.e. only diagonal elements are not replaced by their modulus. For a non-quasimonotone problem we have $J^+ \neq J$ which implies

$$\rho(|G|) > \rho(G) \ ; \tag{3.3}$$

hence the excess of an inclusion is amplified *more violently* than small

perturbations.

This "wrapping" effect (see e.g. Jackson [3]) equally appears in a direct use of the variational equation: To include a solution of

$$e'(t) = J(t)\, e(t) + \dots ,$$

the $\overset{\text{lower}}{\text{upper}}$ bound of e must be used with negative elements of J in the computation of the $\overset{\text{upper}}{\text{lower}}$ bound of e. This corresponds to the use of the 2s×2s-matrix

$$\begin{pmatrix} D+R^{+} & R^{-} \\ R^{-} & D+R^{+} \end{pmatrix}$$

which has the combined spectrum of $J = (D + R^{+} + R^{-})$ and $J^{+} = (D + R^{+} - R^{-}) = (D + |R|)$.

A well-known remedy (cf. e.g. Moore [6]) is the following: Represent the inclusion at each t_μ by $\tilde{E}_\mu = A_\mu \hat{E}_\mu$, $A_\mu \in \mathbb{R}^{s\times s}$. If

$$E_\nu = G_{\nu,\nu-1}\, \tilde{E}_{\nu-1} + C_\nu$$

is the exact local inclusion (see (2.2)), compute

$$A_\nu := G_{\nu,\nu-1}\, A_{\nu-1} , \tag{3.4}$$

$$\hat{E}_\nu := \hat{E}_{\nu-1} + W[A_\nu^{-1} C_\nu] , \tag{3.5}$$

where W is the *wrapping* operation which maps a bounded set in $\mathbb{R}^s$ into the smallest enclosing interval. Now the correct propagation is maintained by (3.4), and each inclusion *increment* C_ν is only wrapped twice (instead of n-ν times), by (3.5) and in the "output" formation

$$\tilde{E}_\nu = W[A_\nu \hat{E}_\nu] . \tag{3.6}$$

(3.6) is needed for the evaluation of various terms in the step proceeding from t_ν. Hereby, C_ν may be distorted by a factor $\|A_\nu\| \|A_\nu^{-1}\| = \mathrm{cond}(A_\nu)$, see (3.5) and (3.6).

Hence, in the use of (3.4)/(3.5) the growth of $\mathrm{cond}(A_\nu)$ has to be monitored and the representation must be restarted if necessary:

$$\tilde{E}_\nu = A_\nu \hat{E}_\nu =: I \hat{\tilde{E}}_\nu .$$

Other tricks to counteract the effect of (3.2) - see e.g. Conradt [1, section 6] - also suffer if the condition of A_ν of (3.4) grows with ν.

None of the presently suggested algorithms is suitable for *stiff problems* (1.1) because polynomial approximations in t are used.

The Accumulated Excess

At some t_ν, we compare the computed inclusion $\tilde{E}_\nu$ to the true error $e(t_\nu)$ (cf. (1.2)) or the true inclusion set $E(t_\nu)$ in case of a set initial condition Y_0 for (1.1) and define

$$\text{global (=accumulated) excess } X_\nu := \tilde{E}_\nu \dot{-} E(t_\nu) \ . \tag{4.1}$$

The computational continuation of the inclusions from $t_{\nu-1}$ to t_ν

- propagates the excess $X_{\nu-1}$ present in $\tilde{E}_{\nu-1}$,
- generates the additional excess $D_\nu =: h_\nu \bar{D}_\nu$, cf. (2.3),
- may introduce further excess by wrapping.

Let us assume that (1.1) is quasimonotone w.r.t. componentwise semiorder so that we may disregard the wrapping effects and that all deviations are sufficiently small so that a perturbation approach is justified. Then we have (cf. "Local Analysis" and "Stability")

$$X_\nu = X_{\nu-1} + h_\nu J_\nu X_{\nu-1} + h_\nu \bar{D}_\nu \ . \tag{4.2}$$

For sufficiently small h_ν, $X_\nu \approx X(t_\nu)$ where

$$X'(t) = J(t)\, X(t) + \bar{D}(t) \ , \quad X(0) = \{0\} \ , \tag{4.3}$$

if we assume that we start the inclusion correctly. From (2.5) we have, with appropriate Λ and Γ,

$$\bar{D}(t) = \Lambda(t)\, h^p + \Gamma(t)\, (\operatorname{diam} \tilde{E}(t))^2 + \text{ small terms } . \tag{4.4}$$

Case 1: Point initial condition for (1.1)

Here $\tilde{E}(t) = e(t) + X(t)$ so that $\operatorname{diam} \tilde{E}(t) = \operatorname{diam} X(t)$. This leads, with (4.4) and (4.3), to

$$X(t;h) = \bar{X}(t)\, h^p + O(h^{p+1}) \tag{4.5}$$

as long as the "small terms" do not become dominant.

This means that the *tightness* of our inclusions (of the error as well as of the true solution) depends on the order p of the algorithm and on the stepsize used, as in usual in o.d.e. algorithms.

Case 2: Interval initial condition Y_0 for (1.1)

$$\begin{aligned} &\operatorname{diam} \tilde{E}(0) = \operatorname{diam} E(0) = \operatorname{diam} Y_0 =: \varepsilon_0 > 0 \ , \\ &\operatorname{diam} \tilde{E}(t) \geq \operatorname{diam} E(t) = \operatorname{diam} Y(t) = r(t)\varepsilon_0 + O(\varepsilon_0^2) \ . \end{aligned} \tag{4.6}$$

Substitution of (4.4)/(4.6) into (4.3) now yields

$$X(t;h,\varepsilon_0) \geq \bar{X}_1(t)h^p + \bar{X}_2(t)\varepsilon_0^2 + \text{higher order terms} \ . \tag{4.7}$$

This means that a reduction of h cannot improve the inclusion beyond the second term. However, this *unavoidable* excess from the quadratic terms in the deviation is $O(\varepsilon_0^2)$ while the error (and solution) tube diameter is $O(\varepsilon_0)$, see (4.6).

The behavior (4.5) and (4.7) is well displayed in numerical computation; results of some experiments are shown in Table 1.

Methods

We can only sketch the two fundamental approaches and must refer to the literature for more detail:

1) Defect Correction: Let $d(t) := \tilde{y}'(t) - f(t,\tilde{y}(t))$ denote the *defect* of $\tilde{y}$. Then we have (cf. (1.2) and (2.2))

$$\begin{aligned} e(t_\nu) &= e(t_{\nu-1}) + \int_{t_{\nu-1}}^{t_\nu} [f(\tau,\tilde{y}(\tau)) - f(\tau,\tilde{y}(\tau)-e(\tau))]d\tau + \int_{t_{\nu-1}}^{t_\nu} d(\tau)\, d\tau \ , \\ E_\nu &\subseteq \tilde{E}_{\nu-1} + \int_{t_{\nu-1}}^{t_\nu} J^+(\tau)E(\tau)\, d\tau + \int_{t_{\nu-1}}^{t_\nu} \text{incl. } \{d(\tau)\}\, d\tau \\ &\quad + \int_{t_{\nu-1}}^{t_\nu} \text{incl. } \{\text{2nd deriv. terms w.r.t. } e\}\, d\tau \end{aligned} \tag{5.1}$$

The solution of the integral inequality (5.1) may be bounded by approximating the resolvent kernel and bounding the remainder. In the last term of (5.1), an a priori estimate for e in $[t_{\nu-1},t_\nu]$ must be used.

This approach was initiated by Schröder (e.g. [8]); an elaborate algorithm has been described by Marcowitz [5] and Conradt [1].

2) Local Expansion: Let y be a truncated Taylor-expansion about $t_{\nu-1}$; denote $y^{(i)}(\bar{t};\bar{t},\bar{y}) =: f_i(\bar{t},\bar{y})$. Then

$$e(t_\nu) = e(t_{\nu-1}) + \sum_{i=1}^{p-1} \frac{h_\nu^i}{i!} [f_i(t_{\nu-1},y_{\nu-1}) - f_i(t_{\nu-1},\tilde{y}_{\nu-1}-e_{\nu-1})] - \frac{h_\nu^p}{p!} f_p(\tau,\bar{y}(\tau))$$

$$E_\nu \subset \tilde{E}_{\nu-1} + \sum_{i=1}^{p-1} \frac{h_\nu^i}{i!} f_i'(t_{\nu-1},\tilde{y}_{\nu-1})\tilde{E}_{\nu-1} + \sum_{i=1}^{p-1} \frac{h_\nu^i}{i!} [f_i'(t_{\nu-1},\tilde{y}_{\nu-1}-\tilde{E}_{\nu-1}) - f_i'(t_{\nu-1},\tilde{y}_{\nu-1})] \tilde{E}_{\nu-1} - \frac{h_\nu^p}{p!} f_p([t_{\nu-1},t_\nu],\bar{Y}_\nu) \qquad (5.2)$$

where $\bar{Y}_\nu$ is an a priori estimate for y in $[t_{\nu-1},t_\nu]$. The approach was initiated by Moore (e.g.[6]); a detailed analysis of an algorithm based upon (5.2) has been presented by Eijgenraam [2].

Obviously, an efficient *implementation* of an inclusion algorithm for (1.2) must rely on a powerful *Computer Algebra system* for the automatic generation of procedures for derivatives and bounds of various kinds, and it must also use an *Interval Arithmetic system* which automatically handles intervals properly (with correct rounding). As both kinds of programming tools are becoming more widely available in standardized forms, the design of transportable and easily usable software for the inclusion problem (1.2) should now become feasible.

References

[1] J.CONRADT, *Ein Intervallverfahren zur Einschließung des Fehlers einer Näherungslösung...*, Freiburger Intervall-Berichte 80/1, 1980.
[2] P.EIJGENRAAM, *The solution of initial value problems using interval arithmetic*, Math. Centre Tracts 144, 1981.
[3] L.W.JACKSON, *Interval arithmetic error-bounding algorithms*, SINUM 12(1975) 223-238.
[4] N.J.LEHMANN, *Fehlerschranken für Näherungslösungen bei Differentialgleichungen*, Numer. Math. 10(1967) 261-288.
[5] U.MARCOWITZ, *Fehlerschätzung bei Anfangswertaufgaben von gew. Diffgln...*, Numer. Math. 24 (1975) 249-275.
[6] R.E.MOORE, *Interval Analysis*, Prentice Hall Inc., 1966.
[7] K.NICKEL, *Using interval methods for the numerical solution of ODEs*, MRC Tech. Summary Rep. #2590, 1982.
[8] J.SCHRÖDER, *Fehlerabschätzung mit Rechenanlagen bei gew. Diffgln. 1. Ordn.*, Numer. Math. 3 (1961) 39-61.
[9] W.WALTER, *Differential- und Integralungleichungen*, Springer-Tracts in Nat. Phil. vol. 2, 1964.

h	$\varepsilon_0 = 0$	$\varepsilon_0 = 2^{-6}$		$\varepsilon_0 = 2^{-4}$		$\varepsilon_0 = 2^{-2}$	
2^{-1}	.25 (-2)	.27 (-2)	17.3	.38 (-2)	6.11		
	27.7	*26.9*		*18.2*			
2^{-2}	.91 (-4)	.10 (-3)	.64	.21 (-3)	.34	.29 (-2)	1.16
	18.7	*11.8*		*3.23*		*2.27*	
2^{-3}	.48 (-5)	.85 (-5)	.054	.65 (-4)	.10	.13 (-2)	.51
	17.4	*2.79*		*1.39*		*1.39*	
2^{-4}	.28 (-6)	.31 (-5)	.020	.47 (-4)	.075	.93 (-3)	.37
	16.7	*1.23*		*1.13*		*1.16*	
2^{-5}	.16 (-7)	.25 (-5)	.016	.41 (-4)	.066	.80 (-3)	.32
	16.4	*1.06*		*1.06*		*1.07*	
2^{-6}	.10 (-8)	.23 (-5)	.015	.39 (-4)	.062	.74 (-3)	.29
	16.2	*1.03*		*1.03*		*1.04*	
2^{-7}	.63 (-9)	.23 (-5)	.015	.38 (-4)	.060	.72 (-3)	.28
	diam $E(\varepsilon_0)$ =	.157 (-3)	*4.0*	.626 (-3)	*4.0*	.253 (-2)	

Table 1. Excess X as a function of h and ε_0.

The problem was $y' = -y^2$, $Y_0 = [1 - \frac{\varepsilon_0}{2}, 1 + \frac{\varepsilon_0}{2}]$, $t \in [0,9]$.

The algorithm used was an implementation of (5.2), with $p = 4$.

The main figures display diam X at $t = 9$, cf. (4.1). The *italic* figures are quotients of their two neighbors. The right-hand figures in the $\varepsilon_0 > 0$ columns are the quotients diam $X(h,\varepsilon_0)$ / diam $E(\varepsilon_0)$.