HOW TO FIGHT THE WRAPPING EFFECT

Karl Nickel

Institut für Angewandte Mathematik
Universität Freiburg

Freiburg i. Br.
West Germany

Abstract: The main purpose of this paper is
not to give Theorems, Algorithms, ...,
but to give insight in the cause and the consequences of the wrapping effect and to derive herefrom indications of how to eliminate it.

Notations: Small letters denote real values, vectors and functions of these. Capital letters denote both real matrices and matrix functions and sets of values, vectors and corresponding functions; in particular intervals of such quantities.

1. The problem

Considered in what follows is the initial value problem for systems of ordinary differential equations

$$u'(t) = f(t,u(t)) \quad \text{for} \quad t \in I := [0,b], \tag{1}$$

$$u(0) = a, \tag{2}$$

where $0 < b \in \mathbb{R}$, $a \in \mathbb{R}^n$, $u: I \to \mathbb{R}^n$, $f: I \times \mathbb{R}^n \to \mathbb{R}^n$. It is always assumed that at last one solution $\hat{u}$ exists in I, where $\hat{u}$ is from an appropriate function space. For simplicity only "exact" solutions are treated; i.e. no approximations, no round off errors, no truncation errors are considered. In order to express the dependance of the solu-

tion $\hat{u}$ with respect to the initial vector a the notation

$$\hat{u}(t;a) \quad \text{for the solutions of} \quad (1), (2)$$

is used.
Wanted are set functions $U(t) \subseteq \mathbb{R}^n$ such that

$$(3) \qquad \hat{u}(t;a) \in U(t) \quad \text{for} \quad t \in I$$

for all solutions $\hat{u}$ of (1), (2). Herein the set U may be an n-dimensional interval, a ball or another suitable set which is easy to determine.

Problem I:

Let $A \subseteq \mathbb{R}^n$ be a bounded set and replace the initial value condition (2) by the initial inclusion condition

$$(2') \qquad u(0) = a \in A.$$

Wanted are again set functions U(t) such that the inclusion (3) holds for all $a \in A$.

Definition: The inclusion (3) is called optimal under the initial inclusion (2') if

$$\forall\, t \in I\ \forall\, y \in U(t)\ \exists\, a \in A\colon y = u(t;a);$$

i.e. if the solutions u(t;a) "fill out completely" the set function U(t) for $a \in A$.

In the Sections 3, 5 and 6 of this paper also the extended

Problem II:

is considered: Let F(t,y) be a bounded set in $\mathbb{R}^n$ for $t \in I$ and $y \in \mathbb{R}^n$ and replace the differential equation (1) by the differential inclusion

$$(1') \qquad u'(t) \in F(t,u(t)) \quad \text{for} \quad t \in I.$$

To be solved is again the inclusion (3) for all solutions of (1'), (2') and special attention is given again to the above defined optimality.

In the paper [6] the author gave a survey on interval methods for the numerical solution of the problem (1), (2). It contains a list of 123 publications from this field. In it also the wrapping effect is regarded. In the meantime a new publication on this effect appeared by Gambill and Skeel [2].

2. Moore's Example

In Figure 1 the well known example of R.E. Moore to the Problem I is sketched. It should be self-explaining and shows that by using intervals U(t) (at the points $t = \pi/6$, $t = \pi/3$, $t = \pi/2$) no optimality can be obtained. This is due to the fact that the set $\{\hat{u}(t;a) \mid a \in A\}$ (a rotating square) can not optimally be wrapped in intervals.

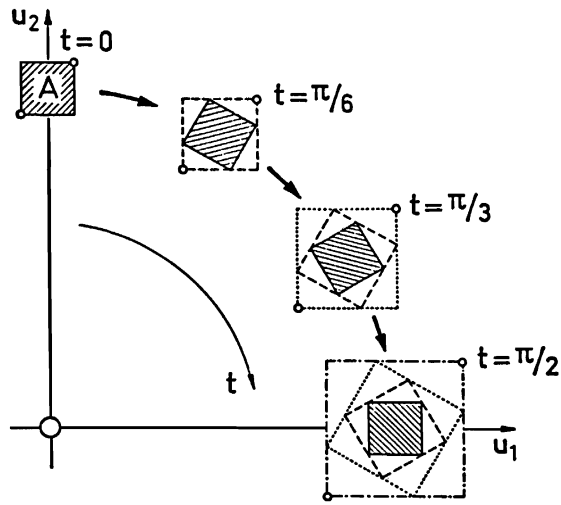

Figure 1. Solution of the differential system

$$\begin{cases} u_1' = u_2, \\ u_2' = -u_1 \end{cases} \tag{4}$$

with the initial data $u(0) = A$

One can show rather easily that after only one revolution ($t = 2\pi$) a blow up of the optimal interval inclusion occurs by a factor of $e^{2\pi} = 535.4 \ldots$ (!). This is due to the use of intervals for U(t). Such a result is most certainly completely intolerable. It occurs, although the system (4) is extremely simple, namely

i) dimension n = 2,

ii) linear system,

iii) homogenous system,

iv) constant coefficients,
i.e. autonomous system.

Hence, in a more general case (n > 2, nonlinear) one expects the worst. Is this true? What is the reason? What can be done? In the following Sections answers to these questions will be given.

3. Systems without wrapping effect

Fortunately, there are large classes of differential equations, where no wrapping effect occurs. A very simple such class is given in what follows.

Definition: Let $f: I\times\mathbb{R}^n \to \mathbb{R}^n$ and denote $f = (f_1, f_2, \cdots, f_n)$ and $f_i = f_i(t, y_1, y_2, \cdots, y_n)$ for $i = 1(1)n$. Then f is called quasiisotone if all components f_i are isotone (monotonically ascending) with respect to all variables $y_1, y_2, \cdots, y_{i-1}, y_{i+1}, \cdots, y_n$; but not necessarily to y_i for $i = 1(1)n$.

Remark: For n = 1 any function f is quasiisotone.

The following Theorem solves Problem I for the class of (in general nonlinear) differential equations (1), where the right hand side f is quasiisotone:

Theorem: Let the function f in the equation (1) be quasiisotone and continuous and consider as sets A and U only intervals $A = [\underline{a}, \overline{a}]$ and $U(t) = [\underline{u}(t), \overline{u}(t)]$. Then no wrapping effect occurs and the inclusion (3) holds with the following weak optimality

$$\underline{u}(t,\underline{a}) \leqq \hat{u}(t,a) \leqq \overline{u}(t,\overline{a})$$

for all solutions $\hat{u}$ of (1) and (2'). Herein the functions $\underline{u}$ and $\overline{u}$ are the (existing) minimal and maximal solutions of (1), (2).

Remark: This Theorem can not be used for Moore's Example of Section 2 since the system (4) is not quasiisotone.

There is a more general Theorem which solves both Problem I and Problem II. It will not be printed here; see [5].

4. Linear Systems. Problem I

In this Section it will be shown that the wrapping effect can be completely explained, understood and avoided if the system (1) is linear. Hence (1) is written in the form

$$u' = Gu + h, \tag{5}$$

where $G: I \to \mathbb{R}^{n\times n}$ and $h: I \to \mathbb{R}^n$. For simplicity it is assumed that $G, h \in C(I)$. Then the problem (5), (2) has exactly one solution $\hat{u} \in C^1(I)$ which can be written as

$$\hat{u}(t;a) = X(t)a + X(t) \int_0^t X^{-1}(s)h(s)ds. \tag{6}$$

Herein the real matrix function

$$X(t) := \exp \int_0^t G(s)ds \tag{7}$$

is given with the given function $G(t)$. It is also the uniquely determined integral basis to the homogenous system (5) under the initial condition

$$X(0) = E \ (= \text{unit matrix}).$$

It is well known that for all $t \in I$ the inverse matrix $X^{-1}(t)$ always exists.

The dependence of the solution $\hat{u}(t;a)$ in formula (6) with respect to the initial values a is obviously linear i.e.

$$\hat{u}(t;\sigma a_1 + \tau a_2) = \sigma\hat{u}(t;a_1) + \tau\hat{u}(t;a_2)$$

for all $\sigma,\tau \in \mathbb{R}$; $a_1, a_2 \in \mathbb{R}^n$.

Hence, the transformation (6) which maps the initial values $a \in \mathbb{R}^n$ for any fixed $t \in I$ into $\hat{u}(t;a) \in \mathbb{R}^n$ is an affine transformation. From this fact follows immediately the

Theorem: If the initial set A in (2') is a

straight line,
simplex,
parallelepiped,
ellipsoid
convex set, etc.,

then the set $\{\hat{u}(t;a) \mid a \in A\}$ belongs to the same corresponding class.

This Theorem explains immediately the negative result of Moore's Example in Section 2: An n-dimensional interval is an n-dimensional rectangle with axis-parallel sides. This will be transformed for $t > 0$ by formula (6) into an n-dimensional parallelepiped. This can, in general, be "wrapped" (included) by an interval only with a certain loss. By computing $\hat{u}(t;a)$ at many steps $0 < t_1 < t_2 < \cdots$ and wrapping it in an interval at each step this loss occurs at any time $t_1, t_2, \cdots$ and may, therefore, multiply and grow exponentially for large values of t. With this insight in the wrapping effect it is obvious to use the following

Countermeasure: For linear systems (5) do not use intervals for U(t) in the inclusion (3). Instead compute the transformation matrix X(t) as defined in (7). Then the set

$$U(t) := X(t)A + X(t) \int_0^t X^{-1}(s)h(s)ds \tag{8}$$

is the optimal inclusion of all solutions $\hat{u}$ to the initial systems (5), (2').

This neat formula (8), unfortunately, does not say how the matrix-function X(t) should be computed. There are several possibilities. One of them is to evaluate all the eigenpairs of the matrix G(t) for each fixed value $t \in I$. From them one could then construct locally an integral basis X(t) to the homogenous system (5). Another such possibility is to integrate the matrix function G(t) and then to compute the

exponential function in the formula (7) by, say, an infinite series. Both methods are, unfortunately, very laborious and time consuming.

There is, fortunately, a much easier way by exploiting the formula (8) directly: If the set A is eigher a simplex or a parallelepiped it is completely determined by n+1 corners. Hence, it is sufficient to solve (5), (2) for those corners a_ν of A for $\nu = 0(1)n$. Then the solutions $\hat{u}(t;a_\nu)$ of those n+1 real problems (5), (2) give the corners of the desired optimal inclusion; see Figure 2 for n = 2 and an initial interval A.

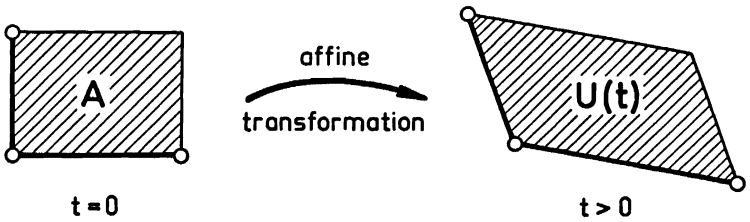

Figure 2. The transformation (8) for n = 2.

This idea has been discovered independently by R. Lohner [3] and by the author [4]. A numerical evaluation has been performed by J. Conradt [1]. It has been applied to the system (4) of Moore's Example of Section 2. His results include numerical integration, round off errors, etc. At t = 2125 (approximately 338 revolutions) his numerical results show a loss of approximately 4 decimal digits. The naive use of an interval inclusion as in Section 2 would, in contrast, have given a loss of 923 (!) decimal digits. What a difference due to the use of intelligence and formula (8) instead of a mindless naive interval computation!

5. Linear Systems. Problem II

In the last Section the following fact has been heavily exploited: The solution $\hat{u}(t;a)$ is a linear function with respect to the initial value a. This can be seen from formula (6). The same formula indicates in addition, that the solution $\hat{u}$ of (5), (2) is also linearly dependent on the function h(t) in (5). Hence, at first sight, it looks as if the idea of Section 4 could also be applied to the differential inclusion

(5') $\quad u' \in Gu + H.$

The difference between (5) and (5') lies in the fact that the real function h(t) has been "blown up" to a set function H(t). Assume, for simplicity, that H(t) is a bounded set in $\mathbb{R}^n$ for any $t \in I$, consisting of continuous function $h \in H$.

This hope and desire is, however, not true, unfortunately. Hence the ideas and methods of the preceding Section 4 can not be carried over to the solution of Problem II. What a pity!

The reason for this unfortunate fact is that the mapping $\int_0^t X(t)X^{-1}(s)h(s)ds$ in formula (6) is in general not an affine mapping. Hence the utilisation of affinity of Section 4 can not be used with respect to the function h in (6).

There is, however, one alternative. Define the continuous function $k: I \to \mathbb{R}^n$ by $k(t) := X^{-1}(t)h(t)$, i.e. let

$$h(t) = X(t)k(t).$$

Then the system (5) is replaced by

(9) $\quad u' = Gu + \dot{X}k$

and the initial value problem (9), (2) has the (obvious) solution (see (6))

$$\hat{u}(t) = X(t)[a + \int_0^t k(s)ds].$$

Kindly note that the matrix function X(t) is defined by formula (7). Hence

X(t) is "known" when G(t) is "given". Since X is always nonsingular the known function h defines an also known function k. Hence the two forms (5) and (9) of a linear system are both equivalent theoretically and from a computational point of view.
There are even real life problems which lead to linear equations of the type (9) instead of (5); where now the functions G,X and k are primarily given. In such a case naturally no transformation from (5) to (9) is needed.

By expanding the function k(t) into a bounded function set K(t) one gets the inclusion

(9') $\quad u' \in Gu + XK$

with the optimal inclusion set function

(10) $\quad U(t) := X(t)[A + \int_0^t K(s)ds]$

to all initial value inclusion problems (9'), (2').

Result:
It should be repeated that the set function U(t) as defined by formula (10) gives an optimal inclusion for both problems I and II.

Remarks:

1) To simplify the computational work the two sets A and K in formula (10) should have the same structure; e.g. they should both be intervals or both be parallelepipeds with parallel sides or ... .

2) The optimal inclusion of solutions of a differential inclusion (9') is possible, because only the function k in equation (9) is expanded into a set while the function G remains a real (matrix) function. To my knowledge no optimal inclusion is known to the solutions of an inclusion of the type (9'), where both functions K and G are set functions.

6. The nonlinear case

In the two previous Sections 4 and 5 the wrapping effect could be controled completely for linear systems. Hence it suggests itself to linearize nonlinear systems and then to treat them by the methods of the Sections 4 and 5.

There are infinitely many possibilities to linearize the system (1) locally by putting

$$f(t,y) = G(t)y + X(t)k(t,y)$$

with a suitable matrix function G(t) and X(t) correspondingly defined by (7). A possible choice of G is

$$G(t) := \frac{\partial f}{\partial y}(t,\tilde{u}(t))$$

with a suitable approximation $\tilde{u}$ to a solution $\hat{u}$ of (1), (2). If f in (1) is continuous then G and k may also be chosen as continuous functions. In this case the initial value problem

$$u'(t) = G(t)u(t) + X(t)k(t,u(t))$$

with the initial condition (2) is obviously equivalent to the Volterra integral equation

$$u(t) = X(t)[a + \int_0^t k(s,u(s))ds].$$

All solutions $\hat{u}$ of this equation can be bounded as in (3) by suitable set functions U and these bounds U can be found and evaluated by standard methods of interval mathematics. These methods also apply if the initial value a and the function k(t,y) are replaced by sets A and K(t,y).

It is not to be expected that in the nonlinear case the wrapping effect can be eliminated completely with this technique. One can expect, however, that the remaining wrapping effect will be "small" if the right hand side f of (1) is only "mildly" nonlinear. If, opposite to this, the function f(t,y) is "strongly" nonlinear with respect to y no method is known to the author to describe and to ban the occuring "nonlinear wrapping effects".

7. Final remarks

The discusssion on the "wrapping effect" of interval methods for the numerical solution of differential equations goes on and on since more than 20 years. Nevertheless one should never lose sight of the following essential facts:

The customary real numerical methods give approximations, not solutions! The error of them is in general not known. It may be very large in particular cases, unknown to the user of such methods.

Compared with this the interval methods have the big advantage to give always guaranteed bounds to the (normally unknown) solutions. Even if they are unfavorable (which may occur with the naive use of interval arithmetic; or without consideration of the wrapping effect) they do give an exact information.

Summary: Any errorbound - even a pessimistic one - is better than no information at all about the error of an approximation. The responsibility of interval research is to find favorable error bounds.

References

[1] Conradt, Jürgen: Ein Intervallverfahren zur Einschließung des Fehlers einer Näherungslösung bei Anfangswertaufgaben für Systeme von gewöhnlichen Differentialgleichungen. Diplomarbeit. Freiburger Intervall-Berichte 80/1. Institut für Angewandte Mathematik, Universität Freiburg i.Br. (1980).

[2] Gambill, Thomas N. and Robert D. Skeel: Logarithmic Reduction of the Wrapping Effect with Applications to Ordinary Differential Equations. University of Illinois, Manuscript (1984).

[3] Lohner, Rudolf: Anfangswertaufgaben im $\mathbb{R}^n$ mit kompakten Mengen für Anfangswerte und Parameter. Diplomarbeit am Institut für Angewandte Mathematik, Universität Karlsruhe (1978).

[4] Nickel, Karl: Bounds for the Set of Solutions of Functional-Differential Equations. MRC Technical Summary Report # 1782, University of Wisconsin, Madison (1977). Annales Polonici Mathematici 42 (1983), 241-257.

[5] Nickel, Karl: Ein Zusammenhang zwischen Aufgaben monotoner Art und Intervall-Mathematik. Numerical Treatment of Differential Equations, Proc. of a conf. held at Oberwolfach, July 4-10, 1976. Ed. by R. Bulirsch, R.D. Grigorieff, and J. Schröder, Springer Verlag, Berlin, Heidelberg, New York, 121-132 (1978).

[6] Nickel, Karl: Using Interval Methods for the Numerical Solution of ODE's. MRC Technical Summary Report # 2590. University of Wisconsin, Madison (1983). Freiburger Intervall-Berichte 83/10. Institut für Angewandte Mathematik, Universität Freiburg i.Br., 13-44 (1983). To appear in ZAMM.