IRREGULARITIES IN THE NUMERICAL TREATMENT
OF NONLINEAR INITIAL VALUE PROBLEMS

R. ANSORGE

Institut für Angewandte Mathematik der Universität Hamburg
D — 2000 Hamburg 13, Federal Republic of Germany

1. Introduction

Let $\mathfrak{B}$ be a Banach space of continuous functions $u(x), x \in R^n$.

A one-parameter family $\{u(t)\} \subset \mathfrak{B}$ $([u(t)](x) = u(x, t))$ (we do not always write down the "space" variables x) is seeked as the solution of the initial value problem

(1) $$u_t = Au + G(t)u, \quad u(0) = u_0, \quad 0 \leqslant t \leqslant T.$$

Here, A is assumed to be a linear differential operator from $\mathfrak{B}_A \subset \mathfrak{B}$ into $\mathfrak{B}$ which does not depend on the "time"-variable t (for convenience).

$G(t)$ $(0 \leqslant t \leqslant T)$ are assumed to be nonlinear operators from $\mathfrak{B}$ into $\mathfrak{B}$. Thus (1) defines an initial value problem with a semilinear partial differential equation (or a system of equations).

For the moment, let the operators $G(t)$ be uniformly and globally Lipschitz-continuous, i.e.:

(2) $$\exists L: \quad \|G(t)u - G(t)v\| \leqslant L\|u - v\|, \quad \forall u, v \in \mathfrak{B}, \forall t \in [0, T].$$

Assume that there are uniquely determined genuine solutions of problem (1) for all $u_0 \in \mathfrak{D} \subset \mathfrak{B}_A$ which we denote by

$$u(t) = E_0(t)u_0, \quad 0 \leqslant t \leqslant T, \quad E_0(t): \mathfrak{D} \to \mathfrak{B}.$$

The operators $E_0(t)$ are called the "solution-operators" of problem (1) on $\mathfrak{D}$.

Obviously, the mapping $[0, T] \xrightarrow[E_0(t)u_0]{} \mathfrak{B}$ is continuous for every fixed $u_0 \in \mathfrak{D}$.

Let us now consider the linear problem belonging to (1), i.e. the problem

(3) $$u_t = Au, \quad u(0) = u_0, \quad 0 \leqslant t \leqslant T,$$

using the same operator A as in (1), which we assume to be properly posed; i.e.,

we assume that there are genuine solutions for all $u_0 \in \mathfrak{F} \underset{\text{dense}}{\subseteq} \mathfrak{B}$ and the solution-operators $E_0^{(\text{lin})}(t)$ of problem (3) are uniformly bounded:

(4) $$\|E_0^{(\text{lin})}(t)\| \leqslant \varkappa_0, \quad \forall t \in [0, T].$$

Following an elementary theorem, there exist certain unique linear and continuous extensions $E^{(\text{lin})}(t)$ of the operators $E_0^{(\text{lin})}(t)$ from the domain $\mathfrak{F}$ to the domain $\mathfrak{B}$:

$$E^{(\text{lin})}(t): \mathfrak{B} \to \mathfrak{B}, \quad \text{linear}, \quad 0 \leqslant t \leqslant T,$$
$$E^{(\text{lin})}(t) = E_0^{(\text{lin})}(t) \quad \text{on } \mathfrak{F},$$
$$\|E^{(\text{lin})}(t)\| \leqslant \varkappa_0.$$

As it was shown by Thompson [1], problem (1) is equivalent to the problem of solving the integral equation

$$u(t) = E^{(\text{lin})}(t)u_0 + \int_0^t E^{(\text{lin})}(t-\tau)G(\tau)u(\tau)d\tau$$

if $u_0 \in \mathfrak{D}$. Writing this with use of the seeked solution-operators $E_0(t)$ of problem (1), we have

(5) $$E_0(t)u_0 = E^{(\text{lin})}(t)u_0 + \int_0^t E^{(\text{lin})}(t-\tau)G(\tau)E_0(\tau)u_0 \, d\tau.$$

Using relations (2) and (4), it immediately follows from (5) that the relation

$$\max_{0 \leqslant t \leqslant T} \|E_0(t)u_0 - E_0(t)v_0\| \leqslant \varkappa_0 \|u_0 - v_0\| + \varkappa_0 LT \max_{0 \leqslant t \leqslant T} \|E_0(t)u_0 - E_0(t)v_0\|$$

holds. Therefore the nonlinear solution-operators $E_0(t)$ are uniformly Lipschitz-continuous on $\mathfrak{D}$, at least for $T < \dfrac{1}{\varkappa_0 L}$, with the Lipschitz-constant $\dfrac{\varkappa_0}{1 - \varkappa_0 LT}$.

Using the same elementary theorem as for the linear problem, it follows from this Lipschitz-continuity that there are continuous extensions for the nonlinear solution operators which are defined uniquely on every domain $\mathfrak{U} \subset \mathfrak{B}$ with $\mathfrak{D} \underset{\text{dense}}{\subseteq} \mathfrak{U}$.

The sets $\{u(t)\}$ which are generated by the operators $E^{(\text{lin})}(t)$ and $E(t)$

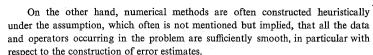$$u(t) = E^{(\text{lin})}(t)u_0, \quad u_0 \in \mathfrak{B},$$
$$u(t) = E(t)u_0, \quad u_0 \in \mathfrak{U}$$

are called "generalized solutions" of problem (3) and problem (1), respectively.

It often happens that not only the genuine solutions but also the generalized solutions are of practical interest because of the fact that the data of a given physical problem often do not fulfil the regularity conditions of an existence theorem belonging to the mathematical formulation of the physical problem.

But also the Lipschitz condition (2) often is not fulfilled, although the existence of a unique solution may be guaranteed if we use an appropriate definition of a "solution".

Therefore one is interested in solving nonlinear problems numerically also in cases of occurring irregularities of the kind just described.

On the other hand, numerical methods are often constructed heuristically under the assumption, which often is not mentioned but implied, that all the data and operators occurring in the problem are sufficiently smooth, in particular with respect to the construction of error estimates.

In this paper the problems arising from this contradiction shall be discussed in the case of solving initial value problems by means of difference methods of the kind

(6) $$\sum_{\nu=0}^{k} A_\nu(h)u_{n+\nu} + h \sum_{\nu=0}^{k} B_\nu(h)G(t_{n+\nu})u_{n+\nu} = 0.$$

This leads to the following questions:

1. Assume that a given method converges (for decreasing step-size) of a certain order towards the exact solution if the initial values are sufficiently smooth (and if the operators $G(t)$ are Lipschitz-continuous). Does the method also converge in the case of less smooth data and (if it converges) what is the order of this convergence?

2. Is it possible to describe the influence of irregularities concerning the operators $G(t)$ under the assumption that the initial values are smooth?

3. What are the consequences if the regularity both of the operators $G(t)$ and of the initial values is affected (combination of question 1 and question 2).

Question 2 (and therefore also question 3) is not yet solved in the case of partial differential equations.

For $A = \Theta$ (the zero-operator) and $\mathfrak{B} = \mathbf{R}$, (1) becomes the ordinary nonlinear first order problem

(7) $$u' = g(t, u), \quad u(0) = u_0, \quad 0 \leqslant t \leqslant T.$$

$(g(t, u) := G(t)u)$ for which question 1 (and therefore also question 3) does not arise.

For linear problems $(G(t) \equiv \Theta)$, question 1 was answered by Peetre and Thomée [2] in 1967 with use of methods of interpolation theory.

It ist well known [3] that a difference scheme which is consistent with a semilinear initial value problem, having Lipschitz-continuous nonlinearities $G(t)$, converges if and only if the corresponding linear scheme (approximating the corresponding linear initial value problem) converges.

Thus, a Lipschitz-continuous nonlinearity does not influence the occurrence of convergence.

It was shown in [4] (by using methods of approximation theory) that also the order of this convergence will not be influenced by the occurrence of a Lipschitz-continuous nonlinear operator $G(t)$. Thus the orders due to Peetre and Thomée are also valid in the Lipschitz-continuous semilinear case.

Therefore we restrict ourselves to the treatment of question 2 in the case of ordinary differential equations.

2. Convergence of difference schemes for ordinary differential equations with non-Lipschitz-continuous right-hand sides

In $\mathfrak{B} = R$ we consider the initial value problem (7). We assume that g is continuous on the compact set

$$D = \{(t, u), 0 \leqslant t \leqslant T, |u| \leqslant b\}.$$

The difference scheme (6) now appears in the form

$$(8) \qquad \sum_{\nu=0}^{k} a_\nu u_{n+\nu} + h \sum_{\nu=0}^{k} b_\nu g(t_{n+\nu}, u_{n+\nu}) = 0,$$

where a_ν, b_ν are real constants ($a_k \neq 0$, $|a_0| + |b_0| > 0$). We assume that the initial set $\{u_0, u_1, \ldots, u_{k-1}\}$ is permissible, i.e.

$$(9) \qquad \lim_{h \to 0} \{u_0, u_1, \ldots, u_{k-1}\} = \{u_0, u_0, \ldots, u_0\}.$$

If one assumes that g is Lipschitz-continuous with respect to u (i.e. the assumptions of the Picard–Lindelöf theorem are fulfilled), the condition of consistency leads to the conditions

$$\sum_{\nu=0}^{k} a_\nu = 0 \quad \text{and} \quad \sum_{\nu=0}^{k} \nu a_\nu + \sum_{\nu=0}^{k} b_\nu = 0,$$

which, after introducing the polynomials

$$\varrho(\lambda) := \sum_{\nu=0}^{k} a_\nu \lambda^\nu, \quad \sigma(\lambda) := \sum_{\nu=0}^{k} b_\nu \lambda^\nu,$$

can be rewritten in the form

$$(10) \qquad \varrho(1) = 0 \quad \text{and} \quad \varrho'(1) + \sigma(1) = 0.$$

The stability condition due to Lax and Richtmyer [5] (formulated for the k-step method (6)) (see also [3]) gives for the special case considered here:

$$(11) \qquad \begin{aligned} &\varrho(\lambda) = 0 \Rightarrow |\lambda| \leqslant 1, \\ &\varrho(\lambda) = 0 \land |\lambda| = 1 \Rightarrow \lambda \text{ is a distinct root of } \varrho. \end{aligned}$$

(10) and (11) are the main conditions of the convergence theorem due to Dahlquist [6], which thus appears as a special case of the Lax–Richtmyer-theory, generalized to semilinear problems.

It should be mentioned that for this special case of difference schemes for the solution of ordinary differential equations, conditions (10), (11) are of purely algebraic structure. Therefore one can look at these conditions independently of regularity assumptions.

Furthermore, let us emphasize the fact that the convergence of the method is not disturbed if there are, besides the root $\lambda_1 = 1$ (which always occurs because of condition (10)), further (distinct) roots of modulus one, provided that the uniqueness of the solution of problem (7) is guaranteed by the Lipschitz condition.

E.g.: In view of the above mentioned Dahlquist-theorem, the well known midpoint rule

$$(12) \qquad u_{n+2} - u_n - 2hg(t_{n+1}, v_{n+1}) = 0$$

$$(k = 2, a_2 = 1, a_1 = 0, a_0 = -1, b_2 = 0, b_1 = -2, b_0 = 0)$$

converges if g is Lipschitz-continuous with respect to u, because condition (10) is fulfilled, and the roots of $\varrho(\lambda)$ are $\lambda_1 = 1$ and $\lambda_2 = -1$.

Considering the problem

$$\dot{u} = g(t, u), \qquad u(0) = 0, \ t \geqslant 0$$

with

$$g(t, u) := \begin{cases} 0 & \text{for} \quad t = 0, \\ 2t & \text{for} \quad u < 0, \\ 2t - 4\dfrac{u}{t} & \text{for} \quad 0 \leqslant u \leqslant t^2 \\ -2t & \text{for} \quad u > t^2 \end{cases} \Bigg\} \ (t > 0),$$

we see that g is a continuous but not Lipschitz-bounded function.

Nevertheless, the solution

$$(13) \qquad u(t) = \tfrac{1}{3} t^2$$

is unique (Hartmann, [7]). But, if one chooses the permissible initial set $u_0 = 0$, $u_1 = -h^2$, the midpoint rule gives

$$u_{2n} = (2nh)^2, \qquad u_{2n+1} = -\big(2(n+1)h\big)^2 \quad (n = 0, 1, 2, \ldots);$$

obviously, for $n \to \infty$, $h \to 0$, $nh \to t$, the numerical solution does not converge to the exact solution (13).

On the other hand, we have the following theorem due to Taubert [8]:

(i) *Assume that conditions* (10), (11) *are fulfilled with the one restriction that* $\lambda_1 = 1$ *is the only root on the unit circle;*

(ii) *assume that the solution of problem* (7) *is unique.*

Then method (8) *is convergent* (and this result also holds if g is not continuous but if (8) has a unique solution in the sense of Filippov [9]) (under the additional and more technical assumption that condition (9) is guaranteed by $\max_{i=1,\ldots,k} |u_i - u_0| = O(h)$).

For the purpose of this paper, it is sufficient to consider functions g which are continuous.

Let us emphasize the fact that also in Taubert's theorem, the uniqueness of the solution is an important assumption. And the example shows that (also for continuous functions g) the convergence may be really impaired if there are more roots on the unit circle than the one (distinct) root $\lambda_1 = 1$ (according to condition (i) of Taubert's theorem).

Because of the fact that there are many customary methods having more roots on the unit circle than the root $\lambda_1 = 1$, the following question arises:

Up to which border between uniqueness (of the solution of problem (7)) (without further information) and uniqueness due to the Lipschitz condition is it allowed to have more distinct roots on the unit circle than the root $\lambda_1 = 1$?

Obviously, this question cannot be answered completely.

Therefore, let us consider some well-known uniqueness theorems (e.g. the theorems due to Nagumo, Perron, Kamke, Coddington–Levinson) and let us ask whether the conditions of these theorems imply the permissibility to have more roots of modulus 1 than the root $\lambda_1 = 1$ or whether they do not.

In this context, let us quote Perron's theorem [10]:

(i) *Assume that the inequality*
$$|g(t, u) - g(t, v)| \leqslant \tilde{h}(t, |u-v|)$$
holds for all (t, u), $(t, v) \in D$, *where* $\tilde{h}(t, \delta)$ *is a continuous function (for all* $t \in [0, T]$ *and for all* $\delta \geqslant 0$*) with the property* $\tilde{h}(t, 0) = 0$.

(ii) *Assume that* $\tilde{\lambda}(t) \equiv 0$ *is the only solution of the problem*
$$\dot{\tilde{\lambda}} = \tilde{h}(t, \tilde{\lambda}), \quad \tilde{\lambda}(0) = 0, \quad 0 \leqslant t < \gamma$$
for all γ *with* $0 < \gamma < T$.

Then the solution of problem (7) is unique.

In Kamke's theorem (which comprehends the Nagumo criterion), $\tilde{h}$ is allowed to be discontinuous at the point $t = 0$ (e.g. Nagumo: $\tilde{h}(t, \delta) = \delta/t$), and in the theorem due to Coddington and Levinson, $\tilde{h}$ has to be continuous with respect to δ for every fixed t and measurable with respect to t for every fixed δ.

It was Olech [11] who showed in 1960 that the theorems of Kamke and Coddington–Levinson are not more general than the Perron criterion:

If the assumptions of Kamke or Coddington–Levinson are fulfilled, one can find a function ω_g fulfilling all the conditions which occur in Perron's theorem with respect to $\tilde{h}$.

This function is given by

(14) $\omega_g(t, \delta) := \sup\limits_{|u-v|=\delta} |g(t, u) - g(t, v)|$.

Indeed, this function is continuous for $0 \leqslant t \leqslant T$, $\delta \geqslant 0$, and $\tilde{\lambda}(t) = 0$ is a (trivial) solution of

(15) $\dot{\tilde{\lambda}} = \omega_g(t, \tilde{\lambda}), \quad \tilde{\lambda}(0) = 0$.

If $\tilde{\lambda}(t) \equiv 0$ is the *only* solution of (15), we will call ω_g an "Olech function".

Therefore, let us now assume that the uniqueness of the solution of the given problem (7) is guaranteed by the property that the function ω_g (constructed from the given function g by (14)) is an Olech function.

ω_g often is a monotone increasing function with respect to the second variable (and then it can be replaced by the function

$$\hat{\omega}_g := \sup\limits_{|u-v|\leqslant\delta} |g(t, u) - g(t, v)|).$$

Furthermore, we have the trivial property

(16) $\forall c = \text{const} > 0 \quad \omega_{cg} = c\omega_g$;

and if ω_g is an Olech function with respect to g, also $c\omega_g$ is an Olech function with respect to cg, provided that $0 < c \leqslant 1$.

On the other hand, it does not necessarily follow from the property of ω_g being an Olech function that also $c\omega_g$ is an Olech function for all $c > 1$ (counterexamples exist). But there are many classes of problems for which this implication is true.

An answer to our question posed above can now be given by the following theorem (see [12]).

THEOREM (Taubert, Hamburg). *Consider the initial value problem*
$$\dot{u} = g(t, u), \quad u(0) = u_0, \quad 0 \leqslant t \leqslant T$$
and assume that

(I) *the function g is continuous on* D,

(II) *the uniqueness of the solution of the problem is guaranteed by the property that* ω_g *is a monotone increasing (with respect to the second variable) Olech function*,

(III) *for all* $\tau > 0$ *also* $\tau\omega_g$ *is an Olech function with respect to* τg,

(IV) *the Dahlquist conditions (10), (11) are fulfilled*,

(V) $|u_i - u_0| = O(h)$ *for* $h \to 0$ ($i = 1, \ldots, k$);

then the difference scheme (8) is convergent (i.e.: also more roots than $\lambda_1 = 1$ *are allowed to be of modulus 1).*

We omit the proof; we just discuss the confining condition (III) which describes the above mentioned border more precisely:

For this purpose, we treat the example
$$g(t, u) := \begin{cases} 0 & \text{for} \quad t = 0, \\ \frac{1}{2}t & \text{for} \quad u < 0, \\ \frac{1}{2}t - \dfrac{u}{t} & \text{for} \quad 0 \leqslant u \leqslant t^2 \\ -\frac{1}{2}t & \text{for} \quad u \geqslant t^2 \end{cases} \Bigg\} \quad t > 0$$

and $u(0) = 0$.

This function g is continuous and the Nagumo criterion is fulfilled. The unique solution of the corresponding initial value problem is
$$u(t) = \tfrac{1}{2}t^2 \quad (\Rightarrow: \text{condition (I) is fulfilled}).$$

The difference scheme

(17) $u_0 = 0, \quad u_1 = h^2, \quad u_{n+2} = u_n + 4hg_{n+1} - 2hg_{n-1}$

fulfils conditions (IV), (V) of Taubert's theorem.

We have
$$\omega_g(t, \delta) = \begin{cases} 0 & \text{for} \quad t = 0, \\ \delta/t & \text{for} \quad 0 \leqslant \delta \leqslant t^2 \\ t & \text{for} \quad \delta > t^2 \end{cases} \Bigg\} \quad t > 0,$$

and ω_g is an Olech function ($\Rightarrow$: condition (II) is fulfilled). However, the problem

$$\dot{\tilde{\lambda}} = \tau\omega_g(t, \tilde{\lambda}), \qquad \tilde{\lambda}(0) = 0$$

is for $\tau = 2$ not only solved by $\tilde{\lambda}(t) \equiv 0$ but also by $\tilde{\lambda}(t) = \alpha t^2$ $(\forall\alpha \in (0, 1])$, and for $\tau > 2$ we have (besides $\tilde{\lambda}(t) \equiv 0$) the solution $\tilde{\lambda}(t) = \frac{\tau}{2} t^2$. Thus condition (III) of Taubert's theorem is not fulfilled.

Indeed, the solution of the difference equation (17) (applied to our example) has the properties

$$u_{2n} \leqslant 0, \qquad u_{2n+1} \geqslant (n+1)^2 h^2,$$

which imply non-convergence!

Condition (III) therefore is an important condition. On the other hand, the conditions of Taubert's theorem are only sufficient for convergence. For certain schemes, condition (III) can be omitted, e.g. for

$$-a_k = a_0 = 1, \qquad a_\nu = 0 \ (\nu = 1, \ldots, k-1), \qquad b_\nu \geqslant 0.$$

References

[1] R. J. Thompson, *Difference approximations for inhomogeneous and quasi-linear equations*, J. Soc. Indust. Appl. Math. 12 (1964), pp. 189–199.

[2] J. Peetre and V. Thomée, *On the rate of convergence for discrete initial-value problems*, Math. Scand. 21 (1967), pp. 159–176.

[3] R. Ansorge, *Konvergenz von Mehrschrittverfahren zur Lösung halblinearer Anfangswertaufgaben*, Numer. Math. 10 (1967), pp. 209–219.

[4] R. Ansorge, C. Geiger und R. Hass, *Existenz und numerische Erfaßbarkeit verallgemeinerter Lösungen halblinearer Angangswertaufgaben*, ZAMM 52 (1972), pp. 597–605.

[5] P. Lax and R. D. Richtmyer, *Survey of the stability of linear finite difference equations*, Comm. Pure Appl. Math. 9 (1956), pp. 167–293.

[6] G. Dahlquist, *Convergence and stability in the numerical integration of ordinary differential equations*, Math. Scand. 4 (1956), pp. 33–53.

[7] P. Hartmann, *Ordinary differential equations*, John Wiley & Sons, New York–London–Sidney 1964.

[8] K. Taubert, *Differenzenverfahren für gewöhnliche Anfangswertaufgaben mit unstetiger rechter Seite*; in: Lecture Notes in Mathem., vol. 395 (editors: R. Ansorge and W. Törnig), pp. 137–148, Springer, Berlin–Heidelberg–New York 1974.

[9] A. F. Filippov, *Differential equations with discontinuous right-hand side*, Amer. Math. Soc. Transl. 42 (1960), pp. 199–231.

[10] O. Perron, *Eine hinreichende Bedingung für die Unität der Lösungen von Differentialgleichungen erster Ordnung*, Math. Z. 28 (1928), pp. 216–219.

[11] C. Olech, *Remarks concerning criteria for uniqueness of solutions of ordinary differential equations*, Bull. Acad. Polon. Sci., Ser. sci. math. astr. et phys. 8 (1960), pp. 661–666.

[12] K. Taubert, *Eine Erweiterung der Theorie von G. Dahlquist*, Computing 17 (1976), pp. 177–185.

О ТОЧНОСТИ СХЕМ ПЕРЕМЕННЫХ НАПРАВЛЕНИЙ ДЛЯ УРАВНЕНИЯ ТЕПЛОПРОВОДНОСТИ В ПРОИЗВОЛЬНОЙ ОБЛАСТИ

А. В. ГУЛИН, И. В. ФРЯЗИНОВ

Институт Прикладной Математики АН СССР, Москва, СССР

В работе установлена сходимость схем Дугласа–Рэчфорда (Д.-Р.) [1] и Писмена–Рэчфорда (П.-Р.) [2] для двумерного уравнения теплопроводности в произвольной области в случае первой краевой задачи в сеточной норме L_2 со скоростями $O(\tau + h^{3/2})$. Здесь $h = \max(h_1, h_2)$, h_α — шаг пространственной сетки в направлении оси координат $0x_\alpha$, $\alpha = 1, 2$, τ — шаг сетки по времени.

Схемы [1], [2] рассматриваются здесь как составные схемы, обладающие свойством суммарной аппроксимации [3]. Устойчивость и сходимость их установлена лишь при предположении неотрицательности соответствующих сеточных операторов. Требование перестановочности операторов не используется. В узлах приграничной зоны используется аппроксимация из [4], [5].

Приводятся оценки скорости сходимости схем Д.-Р. и П.-Р. в случае разрывных коэффициентов, полярных и сферических координат.

1. Постановка задач

1.1. Постановка исходной задачи. Пусть G — ограниченная область в плоскости $0x_1x_2$ с границей Γ, $\bar{G} = G \cup \Gamma$. Предположим, что пересечение области G любой прямой, проходящей через точку $x = (x_1, x_2) \in G$ и параллельной оси координат $0x_\alpha$, состоит лишь из одного интервала $\varDelta_\alpha(x)$. Пусть $Q_T = G \times (0 < t \leqslant T)$, $\bar{Q}_T = \bar{G} \times (0 \leqslant t \leqslant T)$.

Требуется найти непрерывное в $\bar{Q}_T$ решение $u = u(x, t)$ задачи

$$(1.1) \qquad \begin{aligned} \partial u/\partial t &= \varDelta u + f(x, t), \quad (x, t) \in Q_T, \\ u &= v(x, t), \quad x \in \Gamma, \quad 0 < t \leqslant T, \quad u(x, 0) = u^0(x), \quad x \in \bar{G}. \end{aligned}$$

Относительно гладкости входных данных — функций f, v, u^0, а также