Computing 43, 115–131 (1989)

Computing

On the Error Structure of the Implicit Euler Scheme Applied to Stiff Systems of Differential Equations

W. Auzinger*, Wien

Received April 27, 1989

Abstract — Zusammenfassung

On the Error Structure of the Implicit Euler Scheme Applied to Stiff Systems of Differential Equations. In this paper we investigate the structure of the global discretization error of the implicit Euler scheme applied to systems to stiff differential equations, extending earlier work on this subject (cf. [1], [9]). We restrain our considerations to the linear, self-adjoint, constant coefficient case but—in contrast to [1], [9]—we make no assumptions about the nature of the stiff spectrum; the stiff eigenvalues may be arbitrarily distributed on the negative real axis.

Our main result says that the global error of the implicit Euler scheme admits an asymptotic expansion in powers of the stepsize τ which is not asymptotically correct in the conventional sense: Near the initial point $t = 0$ the expansion is spoiled at the $O(\tau^2)$—level by 'irregular' error components which are, however, (algebraically) damped, such that away from $t = 0$ the 'pure' asymptotic expansion reappears. We present numerical experiments confirming this result.

Our considerations should be particularly helpful for a rigorous, quantitative analysis of the structure of the full (space & time) discretization error in the PDE (method of lines) context, and thus for a sound theoretical justification of extrapolation techniques for this important class of stiff problems.

Über die Fehlerstruktur des impliziten Eulerverfahrens bei Anwendung auf steife Differentialgleichungssysteme. In dieser Arbeit wird die Struktur des globalen Diskretisierungsfehlers des impliziten Eulerverfahrens bei Anwendung auf steife Differentialgleichungssysteme untersucht; es handelt sich um eine Weiterführung bestehender Arbeiten zu diesem Thema (siehe [1], [9]). Die vorliegenden Betrachtungen beschränken sich auf den linearen, selbstadjungierten Fall mit konstanter steifer Matrix, jedoch werden—im Gegensatz zu [1], [9]—keine Annahmen über die Struktur des steifen Spektrums getroffen; die steifen Eigenwerte können beliebig auf der negativen reellen Achse verteilt sein.

Unser zentrales Resultat lautet: Der globale Fehler des impliziten Eulerverfahren besitzt eine asymptotische Fehlerentwicklung in Potenzen der Schrittweite τ, die allerdings nicht asymptotisch korrekt im konventionellen Sinn ist: Unmittelbar nach dem Anfangspunkt ist die Entwicklung durch das Auftreten 'irregulärer' Fehlerterme auf $O(\tau^2)$—Niveau gestört. Diese irregulären Komponenten zeigen jedoch ein (algebraisch) abklingendes Verhalten, so daß nach einem gewissen Zeitintervall eine 'reine' asymptotische Entwicklung sichtbar wird. Es werden numerische Resultate präsentiert, die dieses Resultat untermauern.

Die vorliegenden Betrachtungen sollten insbesondere nützlich sein für eine rigorose, quantitative Analyse des vollen Diskretisierungsfehlers (bez. Zeit und Raum) im Zusammenhang mit partiellen Differentialgleichungen (Linienmethode), und damit für eine saubere theoretische Rechtfertigung von Extrapolationstechniken für diese wichtige Klasse von steifen Problemen.

* Institut für Angewandte und Numerische Mathematik, Technische Universität Wien, Wiedner Hauptstrasse 6–10, A-1040 Wien, Austria.

1. Introduction

Within the convergence theory for discretization methods, the concept of an *asymptotic error expansion* in powers of the stepsize τ provides a powerful tool for the theoretical justification of stepsize control mechanisms or acceleration techniques like extrapolation or defect correction. In the present paper we study the existence of an asymptotic error expansion in a *stiff* situation.

It is a remarkable fact that, particularly in the context of stiff problems, the notion of an asymptotic error expansion is often badly misunderstood. To be more precise, let ε_ν denote the global error of some discretization scheme of order p (with stepsize τ) at $t_\nu = \nu\tau$. According to classical results (cf. [8], [12]), ε_ν can be written as

$$\varepsilon_\nu = \sum_{j=p}^{q} \tau^j e_j(t_\nu) + O(\tau^{q+1}) \qquad \text{for} \qquad \tau \to 0 \tag{1.1}$$

with τ-independent functions $e_j(t)$, provided the problem data are sufficiently smooth. However, it is important to realize that for practical values of τ an assertion like (1.1) provides no information at all about the structure of the global error or about its behavior upon grid refinement, *unless* it can be guaranteed that the O-constant is of *moderate size* (compared to the smoothness of the solution sought). This is easy to verify in many ('classical', non-stiff) situations; however, for problems which are well-conditioned but depend on 'critical' problem parameters which are large in size in a way unrelated to the smoothness of the solution, it must be verified that the remainder term of an asymptotic expansion is not influenced by these large parameters.

For stiff problems, it is indeed not obvious whether the O-constant in (1.1) is of moderate size: The $e_j(t)$ are solutions of certain linearized (stiff!) variational problems (cf. section 3) which will usually not be smooth (even if the solution of the original problem is). But the remainder term in (1.1) depends on higher derivatives of the $e_j(t)$, and so it must be expected that the O-constant is not moderate: It may be heavily influenced by the critical problem parameter L (the conventional Lipschitz constant), which is very large in size. Such an expansion would of course be completely useless. Thus, in the stiff case the existence of asymptotic expansions *in the quantitative sense* (i.e., with a moderate O-constant) is a nontrivial question, requiring a careful analysis in the spirit of the B-convergence theory (cf. [6]).

This basic difficulty has been overlooked by many authors, cf. for instance [3], [4], [5], [7], [10]. Therefore the theoretical justification for extrapolation algorithms which is given in all these papers is incorrect.

Recently, a rigorous, quantitative analysis of the error structure of the implicit Euler scheme applied to stiff initial value problems was given in [1] and in [9] (see also [2] for an overview on further results). The analysis in [1] and [9] is based on the assumption that there is only one "cluster" of stiff eigenvalues, which enables the application of singular perturbation techniques. It turns out that the global error contains nonsmooth components which may spoil the error structure immediately after the start (or after a change of stepsize). These irregular error components are, however, exponentially damped, such that the "pure" asymptotic expansion reappears after a short "transient phase."

In the present paper we extend the results of [1] and [9]: We restrain our considerations to the linear, self-adjoint, constant coefficient case, but we make no assumptions about the structure of the stiff spectrum. Several clusters of stiff eigenvalues of different magnitudes or even 'continuously' distributed stiff spectra are admitted. The latter type of stiff problems typically arises in the PDE field (method of lines).

Our main result is stated in Theorem 4.3: The global error of the implicit Euler scheme admits an asymptotic expansion in powers of the stepsize τ which is not asymptotically correct in the conventional sense: Immediately after the initial point $t = 0$ the expansion is spoiled at the $O(\tau^2)$—level by 'irregular' error components. These irregular components, are, however, (algebraically) damped, such that away from $t = 0$ the 'pure' asymptotic expansion reappears.

In section 5 the scope and applicability of our results, especially in the method of lines context, is briefly discussed. Furthermore we present a numerical experiment (τ-extrapolation) confirming the sharpness of our estimates.

2. Preliminaries

We consider a class of linear initial value problems

$$\begin{aligned} u'(t) &= Au(t) + f(t), \qquad 0 \le t \le T \\ u(0) &= u_0 \end{aligned} \tag{2.1}$$

with a constant linear operator A. We are mainly interested in the finite dimensional case, $A \in L(\mathbb{R}^n, \mathbb{R}^n)$, where (2.1) is a system of n ordinary differential equations. (Note, however, that the restriction to a finite dimension is not really necessary; with some technical modifications our considerations may also be understood in a general Hilbert space setting, similarly as in [11].)

Let $\langle \cdot, \cdot \rangle$ denote the Euclidean inner product on $\mathbb{R}^n$; $\|\cdot\|$ will denote the corresponding L_2-norm or its associated operator norm. We assume that A is self-adjoint and negative definite, i.e., the *logarithmic norm of A* satisfies

$$m := \mu_2(A) = \sup_{v \neq 0} \frac{\langle Av, v \rangle}{\langle v, v \rangle} < 0. \tag{2.2}$$

We are interested in the case where (2.1) is *stiff*, that means, $\|A\|$ is very large.

The inhomogeneity $f(t)$ of (2.1) is not required to be of moderate size. We only assume that $A^{-1}f(t)$ is a moderate, smooth function; this is compatible with the existence of smooth solutions to (2.1). Note that in the case that (2.1) represents the semidiscretization in space of an inhomogeneous initial/boundary value problem, the inhomogeneity $f(t)$ will not only represent the source term of the given PDE but will also incorporate the (discretized) boundary conditions; thus $\|f\| \sim \|A\|$ is typical for most practical situations. $\|f\| = O(1)$ is a special case.

Introducing the solution operator

$$(Hg)(t) := \int_0^t e^{(t-s)A} g(s)\, ds \tag{2.3}$$

we may write the solution of (2.1) as

$$u(t) = e^{tA} u_0 + (Hf)(t), \tag{2.4}$$

and the j-th derivatives read[1]

$$u^{(j)}(t) = e^{tA} u_0^{(j)} + (Hf^{(j)})(t) \tag{2.5}$$

with the initial values[2]

$$u_0^{(j)} = A^j u_0 + A^{j-1} f_0 + A^{j-2} f_0' + \cdots + A f_0^{(j-2)} + f_0^{(j-1)}. \tag{2.6}$$

We shall assume that $u(t)$ is *a smooth solution* of (2.1) in the sense that the derivatives up to a certain order are of *moderate size* for $t \in [0, T]$ despite $\|A\|$ large. (2.6) shows that this requires certain compatibility conditions between the initial value u_0 and (the derivatives of) f at $t = 0$. (The case of 'singular', non-compatible initial conditions is briefly discussed in section 5.)

Consider the implicit Euler discretization of (2.1) on a grid $\{t_\nu = \nu\tau, \nu = 0, 1, \ldots\}$ with stepsize τ:

$$\begin{aligned} \frac{1}{\tau}(\eta_\nu - \eta_{\nu-1}) &= A\eta_\nu + f(t_\nu), \\ \eta_0 &= u_0. \end{aligned} \tag{2.7}$$

In the following sections we shall investigate the structure of the global error $\eta_\nu - u(t_\nu)$. In particular, we shall derive an asymptotic expansion in powers of the stepsize τ and present a quantitative estimate for the remainder term of this expansion.

3. The Formal Expansion

Following the work of Gragg [8] (cf. also [1], [12]) one can easily show (using Taylor expansions and equating coefficients of powers of τ) that the global error of the implicit Euler scheme (2.7) applied to (2.1) can be written in the form

$$\eta_\nu - u(t_\nu) = \tau e_1(t_\nu) + \tau^2 e_2(t_\nu) + \cdots + \tau^q e_q(t_\nu) + \rho_\nu \tag{3.1}$$

where the $e_k(t)$ are solutions of the so-called "variational equations", which do not depend on the stepsize τ and are defined in a recursive way:

$$e_1'(t) = Ae_1(t) + f_1(t), \quad f_1(t) := \frac{1}{2} u''(t) \tag{3.2}$$

[1] Note that $(Hg)'(t) = e^{tA} g(0) + (Hg')(t)$.
[2] For any function $g(t)$ we denote $g_0 := g(0)$.

$$e_2'(t) = Ae_2(t) + f_2(t), \quad f_2(t) := \frac{1}{2} e_1''(t) - \frac{1}{6} u'''(t) \tag{3.3}$$

$$\vdots$$

$$e_q'(t) = Ae_q(t) + f_q(t), \quad f_q(t) := \frac{1}{2} e_{q-1}''(t) - \frac{1}{6} e_{q-2}'''(t) + \cdots + \frac{(-1)^{q+1}}{(q+1)!} u^{(q+1)}(t) \tag{3.4}$$

The remainder term ρ_ν is a solution of the difference equation

$$\frac{1}{\tau}(\rho_\nu - \rho_{\nu-1}) = A\rho_\nu + \gamma_\nu, \qquad \gamma_\nu := \frac{1}{\tau} i_{0,\nu} + i_{1,\nu} + \tau i_{2,\nu} + \cdots + \tau^{q-1} i_{q,\nu} \tag{3.5}$$

where the $i_{k,\nu}$ arise from the following Taylor expansions:

$$\begin{aligned} i_{0,\nu} &:= u(t_{\nu-1}) - u(t_\nu) + \tau u'(t_\nu) + \cdots + \frac{(-\tau)^{q+1}}{(q+1)!} u^{(q+1)}(t_\nu) \\ &= \frac{(-\tau)^{q+2}}{(q+1)!} \int_0^1 \sigma^{q+1} u^{(q+2)}(t_{\nu-1} + \sigma\tau)\, d\sigma \end{aligned} \tag{3.6}$$

$$\begin{aligned} i_{1,\nu} &:= e_1(t_{\nu-1}) - e_1(t_\nu) + \tau e_1'(t_\nu) + \cdots - \frac{(-\tau)^q}{q!} e_1^{(q)}(t_\nu) \\ &= \frac{(-\tau)^{q+1}}{q!} \int_0^1 \sigma^q e_1^{(q+1)}(t_{\nu-1} + \sigma\tau)\, d\sigma \end{aligned} \tag{3.7}$$

$$\begin{aligned} i_{2,\nu} &:= e_2(t_{\nu-1}) - e_2(t_\nu) + \tau e_2'(t_\nu) + \cdots - \frac{(-\tau)^{q-1}}{(q-1)!} e_2^{(q-1)}(t_\nu) \\ &= \frac{(-\tau)^q}{(q-1)!} \int_0^1 \sigma^{q-1} e_2^{(q)}(t_{\nu-1} + \sigma\tau)\, d\sigma \end{aligned} \tag{3.8}$$

$$\vdots$$

$$\begin{aligned} i_{q,\nu} &:= e_q(t_{\nu-1}) - e_q(t_\nu) + \tau e_q'(t_\nu) \\ &= \tau^2 \int_0^1 \sigma e_q''(t_{\nu-1} + \sigma\tau)\, d\sigma \end{aligned} \tag{3.9}$$

The general solutions of the variational equations (3.2), (3.3), ... (with arbitrary initial values $e_k(0) = e_{k,0}$) read

$$e_k(t) = e^{tA} e_{k,0} + (Hf_k)(t), \tag{3.10}$$

(cf. (2.3) for the definition of H), and their j-th derivatives are given by

$$e_k^{(j)}(t) = e^{tA} e_{k,0}^{(j)} + (Hf_k^{(j)})(t) \tag{3.11}$$

with the initial values

$$e_{k,0}^{(j)} = A^j e_{k,0} + A^{j-1} f_{k,0} + A^{j-2} f_{k,0}' + \cdots + A f_{k,0}^{(j-2)} + f_{k,0}^{(j-1)}. \tag{3.12}$$

The solution of the remainder equation (3.5) (with initial value ρ_0) is

$$\rho_\nu = (I - \tau A)^{-\nu} \rho_0 + \tau \sum_{l=1}^{\nu} (I - \tau A)^{l-\nu-1} \gamma_l. \tag{3.13}$$

The initial values $e_{k,0}$ and ρ_0 have to be fixed such that

$$0 = \eta_0 - u_0 = \tau e_{1,0} + \tau^2 e_{2,0} + \cdots + \tau^q e_{q,0} + \rho_0; \tag{3.14}$$

thus a natural choice is

$$e_{1,0} = e_{2,0} = \cdots = e_{q,0} = \rho_0 = 0. \tag{3.15}$$

By construction (cf. (3.5), (3.6)–(3.9)) the inhomogeneity γ_ν of the remainder equation (3.13) contains the factor τ^{q+1}. Thus we may use a B-stability argument to estimate ρ_ν by

$$\|\rho_\nu\| \le C\tau^{q+1} \tag{3.16}$$

where C is some τ-independent constant which is influenced by certain derivatives of the $e_k(t)$. So the asymptotic error expansion

$$\eta_\nu - u(t_\nu) = \tau e_1(t_\nu) + \tau^2 e_2(t_\nu) + \cdots + \tau^q e_q(t_\nu) + O(\tau^{q+1}) \tag{3.17}$$

is valid *in the* $\tau \to 0$ *sense*. However, as explained in section 1, an asymptotic relation like (3.17) may be completely useless for realistic values of τ unless it can be guaranteed that the functions $e_k(t)$ and the constant C in (3.16) are of moderate size.

In the present situation it is indeed not obvious whether an expansion (3.17) exists *in the quantitative sense*, i.e. whether the $e_k(t)$ and ρ_ν are independent of $\|A\|$. The main difficulty is caused by the fact that for $e_{k,0} = 0$ the solutions $e_k(t)$ of the variational equations will usually not obey a sufficient number of compatibility conditions at $t = 0$: Due to (3.12) is must be expected that the $e_{k,0}^{(j)} = e_k^{(j)}(0)$ are very large in size because they are influenced by powers of A. The necessary compatibility conditions for the $e_{k,0}$ do not automatically follow from the resp. conditions for u_0.

Thus a detailed investigation is necessary; such an analysis will be presented in section 4 for the case $q = 2$. It will be demonstrated that the remainder term ρ_ν shows a *reduced order* ($O(\tau^2)$) at the first grid points; but it turns out that these order reductions are damped away like $1/\nu$, such that away from $t = 0$ the full order $O(\tau^3)$ reappears.

4. Quantitative Analysis for $q = 2$

Let us consider the case $q = 2$:

$$\eta_\nu - u(t_\nu) = \tau e_1(t_\nu) + \tau^2 e_2(t_\nu) + \rho_\nu \tag{4.1}$$

where $e_1(t)$ and $e_2(t)$ are solutions of (3.2) resp. (3.3).

The solutions of the variational equations.

Let the initial values be fixed by $e_{1,0} = e_{2,0} = 0$ (cf. (3.15)). Then

$$e_1(t) = (Hf_1)(t), \qquad f_1(t) = \frac{1}{2}u''(t) \tag{4.2}$$

which is a moderate (bounded) function due to our smoothness assumptions w.r.t. $u(t)$. The derivatives of $e_1(t)$ are given by (cf. (3.12))

$$\begin{aligned} e_1'(t) &= e^{tA}\underbrace{f_{1,0}}_{e_{1,0}'} + (Hf_1')(t) \\ e_1''(t) &= e^{tA}\underbrace{[Af_{1,0} + f_{1,0}']}_{e_{1,0}''} + (Hf_1'')(t) \\ e_1'''(t) &= e^{tA}\underbrace{[A^2 f_{1,0} + Af_{1,0}' + f_{1,0}'']}_{e_{1,0}'''} + (Hf_1''')(t) \\ &\cdots \end{aligned} \tag{4.3}$$

By assumption on $u(t)$, the $(Hf_1^{(j)})(t)$ are moderate functions; but for $j \ge 2$ the initial values $e_{1,0}^{(j)}$ involve powers of A. Using (2.1) at $t = 0$ we get rid of one factor A and express the $e_{1,0}^{(j)}$ by derivatives of u and f at $t = 0$:

$$\begin{aligned} e_{1,0}' &= \frac{1}{2}u_0'' \\ e_{1,0}'' &= u_0''' - \frac{1}{2}f_0'' = u_0''' - \frac{1}{2}A(A^{-1}f_0'') \\ e_{1,0}''' &= \frac{3}{2}u_0^{IV} - f_0''' - \frac{1}{2}Af_0'' = \frac{3}{2}u_0^{IV} - A(A^{-1}f_0''') - \frac{1}{2}A^2(A^{-1}f_0'') \\ &\cdots \end{aligned} \tag{4.4}$$

Recall that the $A^{-1}f^{(j)}$ are assumed to be of moderate size but not the $f^{(j)}$ (cf. section 2). Moreover, the expression for $e_{1,0}'''$ contains even an explicit factor A that cannot be eliminated on the basis of (2.1). Thus, $e_{1,0}'''$ is even influenced by a factor A^2. Furthermore, higher and higher powers of A will occur within the $e_{1,0}^{(j)}$, $j = 4, 5, \ldots$ In other words: The compatibility conditions that are necessary to guarantee that the higher derivatives of $e_1(t)$ are of moderate size are unrelated to the analogous conditions for the original problem (2.1).

The function $e_2(t)$ depends on $e_1''(t)$:

$$e_2(t) = (Hf_2)(t), \qquad f_2(t) = \frac{1}{2}e_1''(t) - \frac{1}{6}u'''(t). \tag{4.5}$$

This is of moderate size because $u'''(t)$ and $A^{-1}e_1''(t)$ are.[3] The derivatives of $e_2(t)$ are given by (cf. (3.12))

$$\begin{aligned} e_2'(t) &= e^{tA}\underbrace{f_{2,0}}_{e_{2,0}'} + (Hf_2')(t) \\ e_2''(t) &= e^{tA}\underbrace{[Af_{2,0} + f_{2,0}']}_{e_{2,0}''} + (Hf_2'')(t) \\ &\cdots \end{aligned} \tag{4.6}$$

[3] Note that $\|(Hg)(t)\| \le \|\int_0^t e^{(t-s)A}\,ds \cdot A\| \cdot \|A^{-1}g\|$ with $\int_0^t e^{(t-s)A}\,ds \cdot A = e^{tA} - I$.

Using (2.1), (3.2) and the definition of f_1 we can again express the initial values $e_{2,0}^{(j)}$ by derivatives of u and f at $t = 0$:

$$e'_{2,0} = \frac{1}{3}u_0''' - \frac{1}{4}f_0'' = \frac{1}{3}u_0''' - \frac{1}{4}A(A^{-1}f_0'')$$

$$e''_{2,0} = \frac{11}{12}u_0^{IV} - \frac{5}{6}f_0''' - \frac{1}{2}Af_0'' = \frac{11}{12}u_0^{IV} - \frac{5}{6}A(A^{-1}f_0''') - \frac{1}{2}A^2(A^{-1}f_0'') \tag{4.7}$$

...

Again, the occurrence of A-factors is unavoidable. The $e_{2,0}^{(j)}$ behave like $e_{1,0}^{(j+1)}$ and are influenced by higher and higher powers of A.

Moreover, the derivatives of $(Hf_2)(t)$ occurring in (4.6) are influenced by the "critical" derivatives of $e_1(t)$. In particular, $(Hf_2'')(t) = \frac{1}{2}(He_1^{IV})(t) - \frac{1}{6}(Hu^V)(t)$ contains the critical term

$$\int_0^t e^{(t-s)A}e^{sA}\,ds \cdot e_{1,0}^{IV} = te^{tA}e_{1,0}^{IV} \tag{4.8}$$

with

$$e_{1,0}^{IV} = 2u_0^V - \frac{3}{2}f_0^{IV} - Af_0''' - \frac{1}{2}A^2f_0''$$

$$= 2u_0^V - \frac{3}{2}A(A^{-1}f_0^{IV}) - A^2(A^{-1}f_0''') - \frac{1}{2}A^3(A^{-1}f_0''). \tag{4.9}$$

The inhomogeneity of the remainder equation.

For $q = 2$ the inhomogeneity of the remainder equation (3.5) reads

$$\gamma_\nu = \frac{1}{\tau}i_{0,\nu} + i_{1,\nu} + \tau i_{2,\nu} \tag{4.10}$$

Here the $i_{k,\nu}$ are given by (3.6,...):

$$\frac{1}{\tau}i_{0,\nu} = \frac{\tau^3}{6}\int_0^1 \sigma^3 u^{IV}(t_{\nu-1} + \sigma\tau)\,d\sigma = O(\tau^3) \tag{4.11}$$

is valid with a moderate O-constant[4] due to our smoothness assumptions w.r.t. $u(t)$. Furthermore,

$$i_{1,\nu} = -\frac{\tau^3}{2}\int_0^1 \sigma^2 e_1'''(t_{\nu-1} + \sigma\tau)\,d\sigma$$

$$= -\frac{\tau^3}{2}e^{t_{\nu-1}A}\int_0^1 \sigma^2 e^{\sigma\tau A}\,d\sigma \cdot e_{1,0}''' - \frac{\tau^3}{2}\int_0^1 \sigma^2(Hf_1''')(t_{\nu-1} + \sigma\tau)\,d\sigma \tag{4.12}$$

[4] In the following, the symbol $O(\tau^p)$ is always to be understood *in the quantitative sense*, i.e., with a moderate O-constant unaffected by $\|A\|$.

with $e_{1,0}'''$ from (4.4), and

$$\tau i_{2,\nu} = \tau^3\int_0^1 \sigma e_2''(t_{\nu-1} + \sigma\tau)\,d\sigma$$

$$= \tau^3 e^{t_{\nu-1}A}\int_0^1 \sigma e^{\sigma\tau A}\,d\sigma \cdot e_{2,0}'' + \tau^3\int_0^1 \sigma(Hf_2'')(t_{\nu-1} + \sigma\tau)\,d\sigma \tag{4.13}$$

with $e_{2,0}''$ from (4.7).

In (4.12), the term involving $Hf_1''' = \frac{1}{2}Hu^V$ is $O(\tau^3)$. All other terms in (4.12) and (4.13) are influenced by the critical derivatives of $e_1(t)$ and $e_2(t)$ (cf. (4.3)–(4.9)). Let the collection of these critical inhomogeneous terms be denoted by $\hat{\gamma}_\nu$; the $O(\tau^3)$-terms are denoted by $\tilde{\gamma}_\nu$. Furthermore we introduce

$$I_k := \int_0^1 \sigma^k e^{\sigma\tau A}\,d\sigma. \tag{4.14}$$

Thus,

$$\gamma_\nu = \hat{\gamma}_\nu + \tilde{\gamma}_\nu = \hat{\gamma}_\nu + O(\tau^3) \tag{4.15}$$

with

$$\hat{\gamma}_\nu = \tau^3 e^{t_{\nu-1}A}\left[I_1 e_{2,0}'' - \frac{1}{2}I_2 e_{1,0}'''\right] + \tau^3\int_0^1 \sigma(t_{\nu-1} + \sigma\tau)e^{(t_{\nu-1}+\sigma\tau)A}\,d\sigma \cdot \frac{1}{2}e_{1,0}^{IV}$$

$$= \tau^3 e^{t_{\nu-1}A}\left[I_1 e_{2,0}'' - \frac{1}{2}I_2 e_{1,0}'''\right] + \tau^3\left[t_{\nu-1}e^{t_{\nu-1}A}I_1\frac{1}{2}e_{1,0}^{IV} + \tau e^{t_{\nu-1}A}I_2\frac{1}{2}e_{1,0}^{IV}\right] \tag{4.16}$$

($e_{1,0}^{IV}$ from (4.9)).

For later use we note:[5]

$$I_1 = (\tau A)^{-2}[I - (I - \tau A)e^{\tau A}], \tag{4.17}$$

I_1 is invertible. Furthermore,

$$\|I_1\| \le \frac{1}{2}, \qquad \|I_2\| \le \frac{1}{3}, \qquad \|I_1^{-1}I_2\| \le \frac{2}{3}. \tag{4.18}$$

(4.17) is simply a Taylor identity; (4.18) can be shown in an elementary way.

Estimation of the remainder term ρ_ν.

At first we note the following well-known B-stability estimates, which are a consequence of (2.2):

[5] A is invertible by assumption (2.2).

Lemma 4.1 *The estimates*

$$\left\| \tau \sum_{l=1}^{\nu} (I - \tau A)^{\ell-\nu-1} \right\| \leq \frac{\left(\frac{1}{1-\tau m}\right)^{\nu} - 1}{m} \tag{4.19}$$

$$\left\| \tau \sum_{l=1}^{\nu} (I - \tau A)^{\ell-\nu-1} A \right\| \leq 1 \tag{4.20}$$

hold with m from (2.2). □

According to (3.13), (4.15) and (4.16), ρ_ν splits into

$$\rho_\nu = \underbrace{\tau \sum_{l=1}^{\nu} (I - \tau A)^{\ell-\nu-1} \tilde{\gamma}_l}_{\tilde{\rho}_\nu} + \underbrace{\tau \sum_{l=1}^{\nu} (I - \tau A)^{\ell-\nu-1} \hat{\gamma}_l}_{\hat{\rho}_\nu}. \tag{4.21}$$

Since $\tilde{\gamma}_\nu = O(\tau^3)$, $\tilde{\rho}_\nu$ is easily estimated on the basis of (4.19):

$$\|\tilde{\rho}_\nu\| \leq \frac{\left(\frac{1}{1-\tau m}\right)^{\nu} - 1}{m} \max_{l=1(1)\nu} \|\tilde{\gamma}_l\| = O(\tau^3). \tag{4.22}$$

The critical term $\hat{\rho}_\nu$ remains to be investigated. Due to (4.16),

$$\begin{aligned} \hat{\rho}_\nu = {} & \tau^3 \tau \sum_{l=1}^{\nu} (I - \tau A)^{\ell-\nu-1} e^{t_{l-1}A} I_1 \left[e''_{2,0} - \frac{1}{2} I_1^{-1} I_2 e'''_{1,0} \right] \\ & + \tau^3 \tau \sum_{l=1}^{\nu} (I - \tau A)^{\ell-\nu-1} t_{l-1} e^{t_{l-1}A} I_1 \cdot \frac{1}{2} e^{IV}_{1,0} \\ & + \tau^3 \tau \sum_{l=1}^{\nu} (I - \tau A)^{\ell-\nu-1} \tau e^{t_{l-1}A} I_1 \cdot \frac{1}{2} I_1^{-1} I_2 e^{IV}_{1,0}. \end{aligned} \tag{4.23}$$

Recall that the expressions for $e'''_{1,0}$, $e''_{2,0}$ and $e^{IV}_{1,0}$ (cf. (4.4), (4.7) and (4.9)) are affected by different powers A^j, $j = 0, 1, 2, 3$. The contributions of these different terms are now discussed separately:

- Terms with a factor A^1 at most are not critical: They can easily be estimated on the basis of Lemma 4.1 (cf. also (4.18)), resulting in $O(\tau^3)$-contributions to $\hat{\rho}_\nu$.
- Terms with a factor A^2 or even A^3 are critical because estimates like (4.20) are not valid with higher power of A instead of merely A. Thus, the usual B-stability estimates are too weak, and the effect of these critical terms has to be studied carefully.

To this end, the following estimates are essential:

Lemma 4.2 *For $\nu \geq 1$ the estimates*

$$\left\| \tau \sum_{l=1}^{\nu} (I - \tau A)^{\ell-\nu-1} e^{t_{\ell-1}A} I_1 A^2 \right\| \leq \frac{1}{t_\nu} \tag{4.24}$$

$$\left\| \tau \sum_{l=1}^{\nu} (I - \tau A)^{\ell-\nu-1} t_{\ell-1} e^{t_{\ell-1}A} I_1 A^3 \right\| \leq \frac{3}{t_\nu} \tag{4.25}$$

hold independent of the stepsize τ.

Proof: At first we prove (4.24). Due to (4.17),

$$\begin{aligned} \tau \sum_{l=1}^{\nu} (I - \tau A)^{\ell-\nu-1} e^{t_{\ell-1}A} I_1 A^2 &= \tau^{-1} [I - (I - \tau A) e^{\tau A}] (e^{\tau A})^{\nu} \sum_{l=1}^{\nu} (I - \tau A)^{-\ell} (e^{\tau A})^{-\ell} \\ &= \tau^{-1} [(I - \tau A)^{-\nu} - (e^{\tau A})^{\nu}]. \end{aligned} \tag{4.26}$$

Since A is assumed self-adjoint and negative definite, we may use a spectral argument, i.e., (4.24) is equivalent to

$$\sup_{x>0} \alpha_\nu(x) \leq \frac{1}{\nu} \tag{4.27}$$

where

$$\begin{aligned} \alpha_\nu(x) &:= [1 - (1 + x)e^{-x}] e^{-\nu x} \sum_{l=1}^{\nu} [(1 + x)^{-1} e^{x}]^{\ell} \\ &= (1 + x)^{-\nu} - e^{-\nu x} > 0. \end{aligned} \tag{4.28}$$

To prove (4.27), we fix some $x_0 \in (0, \sqrt{2})$ and estimate $\alpha_\nu(x)$ for $x \geq x_0$ and $0 \leq x \leq x_0$ separately. Our particular choice for x_0 will be given below.

- For $0 \leq x \leq x_0$, we make use of the fact that $(1 + x)$ is an approximation for e^x.

$$1 - (1 + x)e^{-x} = x^2 \int_0^1 \sigma e^{-\sigma x} d\sigma \leq \frac{1}{2} x^2; \tag{4.29}$$

furthermore,

$$(1 + x)^{-1} e^{x} \leq \left(1 - \frac{x^2}{2}\right)^{-1} = e^{|\ln(1-(x^2/2))|}, \qquad 0 \leq x < \sqrt{2}. \tag{4.30}$$

Thus,

$$\alpha_\nu(x) \leq \frac{1}{2} x^2 e^{-\nu x} \nu [(1 + x)^{-1} e^{x}]^{\nu} = \frac{1}{2\nu} (\nu x)^2 e^{-\nu x} e^{\nu |\ln(1-(x^2/2))|}. \tag{4.31}$$

Assume that x_0 is chosen such that

$$c_0 := \frac{1}{x_0} \left| \ln\left(1 - \frac{x_0^2}{2}\right) \right| < 1. \tag{4.32}$$

Due to $\left| \ln\left(1 - \frac{x^2}{2}\right) \right| \leq c_0 x$ $(0 \leq x \leq x_0)$ we obtain the estimate[6]

[6] Note that $\xi^k e^{-\xi} \leq k^k e^{-k}$ for $\xi \geq 0$, $k \geq 1$.

$$\alpha_\nu(x) \le \frac{1}{2\nu}(\nu x)^2 e^{-\nu(1-c_0)x} = \frac{1}{2\nu(1-c_0)^2}[\nu(1-c_0)x]^2 e^{-\nu(1-c_0)x} \le \frac{2}{e^2(1-c_0)^2}\frac{1}{\nu}. \tag{4.33}$$

- For $x \ge x_0$, $\alpha_\nu(x)$ is exponentially damped with increasing ν:

$$\alpha_\nu(x) < (1+x)^{-\nu} \le (1+x_0)^{-\nu}. \tag{4.34}$$

Now a good choice is $x_0 := 0.75$; we obtain $c_0 \approx 0.44$, $2/(e^2(1-c_0)^2) < 0.87 < 1$ (cf. (4.33)) and $(1+x_0)^{-\nu} < 1/\nu$, $\nu \ge 1$ (cf. (4.34)). This implies (4.27); hence (4.24) holds.

Now we prove (4.25). Due to (4.17) we obtain after some manipulation

$$\begin{aligned}
&\tau \sum_{l=1}^{\nu} (I-\tau A)^{l-\nu-1} t_{l-1} e^{t_{l-1}A} I_1 A^3 \\
&= \tau^{-1}(\tau A)[I-(I-\tau A)e^{\tau A}](e^{\tau A})^\nu \sum_{l=1}^{\nu} (\nu-l)(I-\tau A)^{-l}(e^{\tau A})^{-l} \\
&= \tau^{-1}(\tau A)e^{\tau A}[[(I-\tau A)^{-(\nu-1)} - (e^{\tau A})^{\nu-1}][I-(I-\tau A)e^{\tau A}]^{-1} - (\nu-1)(e^{\tau A})^{\nu-1}].
\end{aligned} \tag{4.35}$$

Thus (4.25) is equivalent to

$$\sup_{x>0} \beta_\nu(x) \le \frac{3}{\nu} \tag{4.36}$$

where

$$\begin{aligned}
\beta_\nu(x) &:= x[1-(1+x)e^{-x}]e^{-\nu x} \sum_{l=1}^{\nu} (\nu-l)[(1+x)^{-1}e^x]^l \\
&= xe^{-x}[[(1+x)^{-(\nu-1)} - e^{-(\nu-1)x}][1-(1+x)e^{-x}]^{-1} - (\nu-1)e^{-(\nu-1)x}] > 0.
\end{aligned} \tag{4.37}$$

Estimating $\beta_\nu(x)$ in a very similar way as $\alpha_\nu(x)$ above yields:

- For $0 \le x \le x_0$,

$$\beta_\nu(x) \le \frac{1}{2\nu}(\nu x)^3 e^{-\nu x} e^{\nu|\ln(1-(x^2/2))|} \le \frac{27}{2e^3(1-c_0)^3}\frac{1}{\nu} \tag{4.38}$$

(c_0 from (4.32)).

- For $x \ge x_0$,

$$\beta_\nu(x) \le \frac{(1+x_0)^{-(\nu-1)}}{e[1-(1+x_0)e^{-x_0}]}. \tag{4.39}$$

Now a good choice is $x_0 := 0.685$; we obtain $c_0 \approx 0.39$, $27/(2e^3(1-c_0)^3) < 2.97 < 3$ (cf. (4.38)) and $e^{-1}[1-(1+x_0)e^{-x_0}]^{-1}(1+x_0)^{-(\nu-1)} < 3/\nu$, $\nu \ge 1$ (cf. (4.39)). This implies (4.36); hence (4.25) holds. □

Remark. (4.26) shows that the estimate (4.24) is equivalent to

$$\|(I-\tau A)^{-\nu} - (e^{\tau A})^\nu\| < \frac{1}{\nu}. \tag{4.40}$$

Estimates of this type are well-known in the theory of time-discretizations for parabolic problems (cf. for instance [11]). However, to our knowledge they have not been used so far in the context of asymptotic error expansions.

Let us now continue the discussion of ρ_ν. Lemma 4.2 enables the estimation of those critical components of $\hat{\rho}_\nu$ (cf. (4.23)) that are influenced by higher powers of A:

- $e^{III}_{1,0}$ and $e^{II}_{2,0}$ are affected by a factor A^2 (cf. (4.4), (4.7)). Due to Lemma 4.2, (4.24), the respective components of $\hat{\rho}_\nu$ can immediately be estimated by

$$\frac{1}{t_\nu} \cdot O(\tau^3). \tag{4.41}$$

- $e^{IV}_{1,0}$ is affected by A^2 and A^3 (cf. (4.9)):
  - The 'A^2-terms' originating from $e^{IV}_{1,0}$ are easily seen to be $O(\tau^3)$; this is a simple consequence of Lemma 4.2. Note, in particular, that the estimate

$$\left\| \tau \sum_{l=1}^{\nu} (I-\tau A)^{l-\nu-1} t_{l-1} e^{t_{l-1}A} I_1 A^2 \right\| < 1 \tag{4.42}$$

    follows easily from (4.24).
  - The 'A^3-terms' originating from $e^{IV}_{1,0}$ can easily be estimated by

$$\frac{3}{t_\nu} \cdot O(\tau^3) \tag{4.43}$$

    on the basis of Lemma 4.2, (4.25).

The O-constants in all the above estimates can be expressed in terms of derivatives of u and $A^{-1}f$.

We end up with

Theorem 4.3 *The global error of the implicit Euler scheme applied to* (2.1) *admits an asymptotic expansion*

$$\eta_\nu - u(t_\nu) = \tau e_1(t_\nu) + \tau^2 e_2(t_\nu) + \rho_\nu \tag{4.44}$$

where $e_1(t)$ and $e_2(t)$ are moderate, τ-independent solutions of the variational equations (3.2), (3.3); *the remainder term ρ_ν can be estimated by*

$$\|\rho_\nu\| \le \left(C_0 + \frac{C_1}{t_\nu}\right)\tau^3 \tag{4.45}$$

with moderate bounds $C_i = C_i(t_\nu)$.

5. Discussion and Numerical Illustration

Theorem 4.3 shows that the error structure of the implicit Euler scheme applied to a stiff system (2.1) is 'almost perfect': For $q = 2$ the global error admits an asymptotic expansion with a $O(\tau^3)$-remainder term ρ_ν which we expect to show a weakly singular behavior (like $1/t_\nu$) near $t = 0$. Concerning longer expansions, the analysis

of section 4 suggests that for $q = 3$ the $O(\tau^4)$-remainder term behaves like $1/t_v^2$ near $t = 0$, and so on. In this paper we do not, however, attempt to enter into the technical details of the general case $q > 2$. Rather, we shall now present a simple numerical experiment in order to illustrate the sharpness of our results.

We consider the 'Prothero-Robinson type' system

$$\begin{aligned} u'(t) &= A(u(t) - g(t)) + g'(t), \\ u(0) &= g(0) \end{aligned} \tag{5.1}$$

with a stiff matrix A arising from the standard space discretization of the 1D-heat equation on a uniform mesh $\{x_i = ih, i = 1, \ldots, N-1\}$ with mesh size $h = 1/N$:

$$A = \frac{1}{h^2}\begin{pmatrix} -2 & 1 & & & & \\ 1 & -2 & 1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \\ & & & & 1 & -2 \end{pmatrix} \tag{5.2}$$

Except for special choices of g, the inhomogeneity of (5.1) is large in size: $\|-Ag + g'\| \sim \|A\| \sim 1/h^2$. The true solution is simply $u(t) = g(t)$. For the present purpose we consider (5.1) as the given, original problem without reference to an underlying PDE problem. This point of view is somewhat artificial but simply serves to suppress the additional effects which are caused in practice by the space discretization process (cf. the remark at the end of this section).

The following numerical results for τ-extrapolation based on the implicit Euler scheme were obtained on a CDC Cyber 180/860 in double precision (≈ 29 decimal digits). $g(t)$ is chosen as the restriction of the smooth function $e^{-t}\cos x$ to the space grid with mesh size $h = 1/128$ ($\|A\| \approx 10^5$).

Table 5.1 displays the global errors (measured in the L_2 norm) of the extrapolated Euler scheme at $t = 0.01$.

error at $t = 0.01$ $h = 1/128$ Table 5.1

tau	Euler	1st EX	2nd EX	3rd EX	4th EX
1/200	1.833E-05				
1/400	9.295E-06	3.168E-07			
1/800	4.682E-06	8.552E-08	1.014E-08		
1/1600	2.350E-06	2.230E-08	1.498E-09	3.146E-10	
1/3200	1.177E-06	5.702E-09	2.055E-10	2.679E-11	9.821E-12
observed order					
	0.98				
	0.99	1.89			
	0.99	1.94	2.76		
	1.00	1.97	2.87	3.55	

error at $t = 0.1$ (global extrapolation) $h = 1/128$ Table 5.2

tau	Euler	1st EX	2nd EX	3rd EX	4th EX
1/200	1.143E-04				
1/400	5.759E-05	8.472E-07			
1/800	2.890E-05	2.151E-07	4.496-09		
1/1600	1.448E-05	5.422E-08	5.744E-10	1.526E-11	
1/3200	7.246E-06	1.361E-08	7.260E-11	9.861E-13	4.033E-14
observed order					
	0.99				
	0.99	1.98			
	1.00	1.99	2.97		
	1.00	1.99	2.98	3.95	

error at $t = 1.0$ (global extrapolation) $h = 1/128$ Table 5.3

tau	Euler	1st EX	2nd EX	3rd EX	4th EX
1/200	8.023E-05				
1/400	4.007E-05	7.662E-08			
1/800	2.003E-05	1.916E-08	2.212E-12		
1/1600	1.001E-05	4.789E-09	1.793E-13	1.351E-13	
1/3200	5.005E-06	1.197E-09	1.791E-14	8.419E-15	2.783E-17
observed order					
	1.00				
	1.00	2.00			
	1.00	2.00	3.62		
	1.00	2.00	3.32	4.00	

Extrapolation works quite well; however, the observed order is a little bit below the optimal value and the increase in accuracy (j-th compared with $(j-1)$-th extrapolation step) apparently decreases with increasing j.

Table 5.2 and 5.3 show the situation at $t = 0.1$ resp. $t = 1.0$ after performing global extrapolation. It is clearly seen that the level of accuracy achieved by the higher extrapolation steps improves significantly with increasing t, and the full quantitative order appears as soon as the algebraic singularity that influences the remainder term of the asymptotic expansion becomes insignificant.

Table 5.4 shows the results at $t = 1.0$ after performing *local* extrapolation. Here the improved performance of extrapolation compared with $t = 0.01$ is not visible. This can easily be explained on the basis of our theoretical results: Local extrapolation means that the integration is *restarted* after each integration interval from the most accurate extrapolated value. It must therefore be expected that the weakly singular behavior of the remainder term also reappears at the beginning of each extrapola-

error at $t = 1.0$ (local extrapolation) $h = 1/128$ Table 5.4

tau	Euler	1st EX	2nd EX	3rd EX	4th EX
1/200	6.811E-06				
1/400	3.454E-06	1.177E-07			
1/800	1.740E-06	3.178E-08	3.768E-09		
1/1600	8.730E-07	8.289E-09	5.577E-10	1.178E-10	
1/3200	4.373E-07	2.119E-09	7.735E-11	1.082E-11	4.293E-12
observed order					
	0.98				
	0.99	1.89			
	0.99	1.94	2.76		
	1.00	1.97	2.85	3.45	

tion interval; thus the damping effect observed for global extrapolation (Table 5.3) cannot take place. Thus, global extrapolation is to be preferred in practice, at least for higher accuracy requirements.

Once more it must be emphasized that the satisfactory performance of extrapolation (especially in its global version) is a consequence of our results of section 4 but cannot be explained on the basis of conventional asymptotic expansions in the (non-quantitative) $\tau \to 0$ sense (cf. the introductory discussion of section 1).

In this paper we have assumed that $u(t)$ is a smooth solution of the given stiff system (2.1) (also the above numerical experiment refers to such a situation). An analysis for the general case of incompatible initial data—such that $u(t)$ is not smooth near $t = 0$—will be more technical but should not be fundamentally different.

The major motivation for the present work is to provide a sound theoretical justification of extrapolation techniques for PDE problems (method of lines approach). Note, however, that in the PDE context the (semi-discrete) stiff system is of an auxiliary nature; the analysis has to be centered about the original PDE (initial/boundary value) problem, and the effects of the space discretization must also be understood. This requires a modified analysis; some work in this direction is under preparation. Furthermore, symmetric schemes (τ^2-extrapolation!) and splitting methods (locally one-dimensional or alternating-direction implicit schemes), which are very important in practice, remain to be investigated.

References

[1] W. Auzinger, R. Frank, F. Macsek, Asymptotic error expansions for stiff equations: The implicit Euler scheme, to appear in SIAM J. Numer. Anal. 27, 1990.
[2] W. Auzinger, R. Frank, Asymptotic expansions of the global discretization error for stiff problems, SIAM J. Sci. Stat. Comput. *10*, 950–963 (1989).
[3] G. Bader, P. Deuflhard, A semi-implicit midpoint rule for stiff systems of ordinary differential equations, Numer. Math. *41*, 373–398 (1983).

[4] P. Deuflhard, Recent progress in extrapolation methods for ordinary differential equations, SIAM Review *27*, 505–535 (1985).
[5] G. Fairweather, J. P. Johnson, On the extrapolation of Galerkin methods for parabolic problems, Numer. Math. *23*, 269–287 (1973).
[6] R. Frank, J. Schneid, C. W. Ueberhueber, The concept of B-convergence, SIAM J. Numer. Anal. *18*, 753–780 (1981).
[7] A. R. Gourlay, J. Ll. Morris, The extrapolation of first order methods for parabolic partial differential equations II, SIAM J. Numer. Anal. *17*, 641–655 (1980).
[8] W. B. Gragg, Repeated extrapolation to the limit in the numerical solution of ordinary differential equations, Ph.D. Thesis, UCLA 1963.
[9] E. Hairer, Ch. Lubich, Extrapolation at stiff differential equations, Numer. Math. *52*, 377–400 (1988).
[10] J. D. Lawson, J. Ll. Morris, The extrapolation of first order methods for parabolic partial differential equations I, SIAM J. Numer. Anal. *15*, 1212–1224 (1978).
[11] M.-N. Le Roux, Semidiscretization in time for parabolic problems, Math. Comp. *33*, 919–931 (1979).
[12] H. J. Stetter, Asymptotic expansions for the error of discretization algorithms for non-linear functional equations, Numer. Math. *7*, 18–31 (1965).

W. Auzinger
Institut für Angewandte und Numerische Mathematik
TU-Wien
Wiedner Hauptstrasse 6–10
A-1040 Wien
Austria