Peter Deuflhard

Uniqueness Theorems for Stiff ODE Initial Value Problems

Preprint SC-87-3, (Juli 1987)

Konrad-Zuse-Zentrum für Informationstechnik;
Heilbronner Straße 10; D-1000 Berlin 31
Peter Deuflhard

Uniqueness Theorems for Stiff ODE Initial Value Problems

Abstract

The paper presents a new uniqueness theory for ODE initial value problems, derived in view of numerical stiff integration. The theory supplies stepsize bounds for stiff integrators that can easily be estimated in extrapolation methods. The additional devices lead to a significant speed-up of computations - in particular in combustion PDE problems.

Keywords: ODE theory, stiff integration, extrapolation methods.

Subject Classification: AMS(MOS): 65 L 05; CR: 5.17
## Contents

0 Introduction .................................................. 1  
1 Preliminary Considerations ................................. 2  
2 Newton-Type Uniqueness Theorems ......................... 5  
3 Stepsize Bounds for Implicit Discretization Methods .... 8  
4 Computational Estimation of Stepsize Bounds in Stiff Extrapolation Integrators .... 12  
5 Numerical Experiments .................................... 17  
Conclusion ..................................................... 22  
References ..................................................... 23
0. Introduction

Reliable numerical algorithms are, in one way or the other, appropriate implementations of uniqueness theorems of the underlying analytic problem. In non-stiff numerical integration, the associated uniqueness theorem for ODE initial value problems is the well-known Picard-Lindelöf theorem - which characterizes the growth of the solution by means of the Lipschitz constant of the right-hand side of the ODE system. For stiff integration, this characterization is known to be inappropriate, but an associated analytic uniqueness theorem in terms of a different characterization seems to be missing - despite of the enormous amount of literature dealing with stiff integration.

It is the purpose of the present paper to fill this gap. The idea behind is, of course, that such an analytic uniqueness theorem should serve as a frame for theoretical investigations of discretization methods for stiff ODE systems. After some preliminary considerations (section 1) two variants of the intended theorem are derived (in section 2) on the basis of an affine invariant convergence theorem due to [10]. The emerging theoretical characterization is then used (in section 3) to determine stepsize bounds for implicit one-step methods. In section 4, a means of estimating the theoretical characteristics directly within semi-implicit extrapolation codes is worked out. Finally, in section 5, numerical experiments (including large scale real life ODE and PDE problems) are presented to demonstrate the efficiency of the additional devices that have been suggested on the new theoretical basis.
1. Preliminary Considerations

Usually, an initial value problem (IVP) for ordinary differential equations (ODE's) is given in the form

\[ y' = f(y), \quad y(0) = y_0 . \]  

(1.1)

For the subsequent proofs, however, the equivalent formulation in terms of a homotopy is preferable:

\[ F(y, \tau) := y(\tau) - y_0 - \int_0^\tau f(y(t)) \, dt = 0 . \]  

(1.2)

Herein the interval length \( \tau \geq 0 \) represents the embedding parameter. Let \( \Gamma \) denote some neighborhood of the graph of a solution of (1.1). Then Peano's existence theorem requires that

\[ L_0 := \sup_{\tau} ||f(y)|| < \infty . \]  

(1.3)

In order to prove uniqueness, the standard approach is to construct the so-called Picard iteration

\[ y^{i+1}(\tau) = y_0 + \int_0^\tau f(y'(t)) \, dt \]  

(1.4)

to be started with \( y^0(t) \equiv y_0 \). From this fixed point iteration, one immediately derives

\[ ||y^{i+1}(\tau) - y^i(\tau)|| \leq \int_0^\tau ||f(y'(t)) - f(y^{i-1}(t))|| \, dt . \]  

(1.5)

Hence, in order to study contraction, the most natural theoretical characterization is in terms of the Lipschitz constant \( L_1 \) defined by

\[ ||f(u) - f(v)|| \leq L_1 ||u - v|| . \]  

(1.6)

With this definition, the sequence \( \{y^i\} \) can be shown to converge to some solution \( y^* \) such that

\[ ||y^*(\tau) - y_0|| \leq L_0 \varphi(L_1 \tau) \]  

(1.7.a)

with

\[ \varphi(s) := \begin{cases} (\exp(s) - 1)/s & s \neq 0 \\ 1 & s = 0 \end{cases} . \]  

(1.7.b)

Moreover, \( y^* \) is unique in \( \Gamma \). This is the main result of the well-known Picard-Lindelöf theorem.
A similar term arises in the analysis of one-step discretization methods for ODE-IVP’s. Let $p \geq 1$ denote the consistency order of such a method, then the discretization error between the analytic solution $y$ and the discrete solution $y_h$ can be represented in the form (see, for example, the recent textbook [17]):

$$\|y_h(\tau) - y(\tau)\| \leq C_p \cdot h^p \cdot \tau \cdot \varphi(L_1 \tau) \quad (1.8)$$

Herein $C_p$ typically is a bound of a higher order derivative of $f$. In order to bound the discretization error, a condition of the kind

$$L_1 \tau \leq C, \quad C = 0(1) \quad (1.9)$$

is needed. Therefore, this characterization is only appropriate for non-stiff discretization methods.

In the beginning of the study of stiff integration, it was first thought that the use of implicit discretization methods would be the essential item to overcome the observed difficulties - see, for instance, the early pioneering paper by Dahlquist [4]. In the next stage of insight, it was recognized that the solution method for the thus arising algebraic equations is equally important: the early paper of Liniger/Willoughby [18] pointed out that fixed point iteration only based on $f$-evaluations for the algebraic equations would once more bring in condition (1.9), whereas a Newton-like iteration were just the method of choice. Much later, so-called semi-implicit or linearly-implicit discretization methods (e.g. Rosenbrock methods, W-methods, extrapolation methods) were constructed that only apply 1 Newton-like iteration per discretization step. Therefore, at the present stage of the development, the essence of non-stiff integration seems to be that only $f$ is sampled, whereas stiff integration additionally requires sampling of the Jacobian $f_y$ or an approximation of it.

With these preparations the natural approach towards the intended uniqueness theorem seems to be replacing the Picard iteration (1.4) by a Newton iteration in function space. For the ordinary Newton method one has

$$F_y(y^i) \Delta y^i = -F(y^i)$$

and

$$y^{i+1} = y^i + \Delta y^i \quad (1.10)$$

or, in more explicit notation:

$$\Delta y^i(\tau) - \int_0^\tau f_y(y^i(t)) \Delta y^i(t) dt =$$

$$= - \left[ y^i(\tau) - y_0 - \int_0^\tau f(y^i(t)) dt \right]. \quad (1.10')$$

Obviously, the above iteration requires global information in terms of $f_y$ rather than just pointwise information as in numerical stiff integration.
Therefore, the simplified Newton method will be the method of choice: one just replaces

\[ F^\prime(y^i) \rightarrow F^\prime(y^0), \quad y^0(t) \equiv y_0 \]

or, equivalently,

\[ f^i(y^i(t)) \rightarrow f^i(y_0) =: A . \]  (1.11)

Insertion of (1.11) into (1.10') then leads to

\[ y^{i+1}(t) - A \int_0^t y^{i+1}(s) \, ds = y_0 + \int_0^t [f(y(t)) - Ay(t)] \, ds \]  (1.12)

Upon comparing (1.12) with (1.4), one may recognize that (1.12) is a Picard iteration for the equivalent ODE

\[ y' - Ay = \bar{f}(y) := f(y) - Ay, \quad y(0) = y_0 \]  (1.13)

Starting with (1.13), a so-called deflated Lipschitz constant has been introduced in [2,7]:

\[ \| \bar{f}(u) - \bar{f}(v) \| \leq L_1 \| u - v \| \]  (1.14)

It is clear that this characterization will lead to the analogon of (1.9), namely

\[ L_1 \tau \leq C, \quad C = O(1) . \]  (1.15)

On the other hand, the definition of \( L_1 \) contains some additional \( r \)-dependence, compare (3.7') in [7]. Hence, condition (1.15) is theoretically unsatisfactory. The alternative is to recall the derivation in terms of Newton's method.
2. Newton-Type Uniqueness Theorems

In this section, the convergence of the Newton-like iteration (1.12) will be used to prove two uniqueness theorems for ODE-IVP's - one for exact and one for approximate Jacobian. Upon applying the variation of constants, one may rewrite (1.12) as

\[ \Delta y'(t) = \int_{t_0}^{t} \exp(A(t-t_0)) [f(y'(t)) - \frac{d}{dt} y'(t)] \, dt \quad (2.1) \]

where \( \exp(At) \) denotes the matrix exponential characterizing the solution of (1.1) for \( f = Ay \). Throughout this section, the main tool of proof will be Theorem 3 of Deuflhard/Heindl [10], which is an affine invariant version of a convergence theorem for Newton-like methods due to Rheinboldt [21]. The results to be derived, however, can only be gained by the affine invariant form of the theorem. Moreover, to replace \( L_1 \), the logarithmic norm \( \mu \) of the Jacobian approximation \( A \), as introduced in [5], will be applied.

Theorem 1 With the notation above let \( f \in C^1(D), \, D \subseteq \mathbb{R}^n \). For the Jacobian \( A := f_y(y_0) \) assume a one-sided Lipschitz condition of the form

\[ \langle u, Au \rangle \leq \mu(u,u) \equiv \mu \|u\|^2 \quad , \quad (2.2.a) \]

where \( \langle \cdot, \cdot \rangle \) denotes an inner product in \( \mathbb{R}^n \), which induces the \( \mathbb{R}^n \)-norm \( \| \cdot \| \). In this norm, assume that

\[ \|f(y)\| \leq L_0 \quad \forall y \in D \quad (2.2.b) \]

\[ \|f_y(u) - f_y(v)\| \leq L_2 \|u - v\| \quad \forall u, v \in D \quad (2.2.c) \]

Then, for \( D \) sufficiently large, existence and uniqueness of the solution of the ODE-IVP (1.1) is guaranteed for

\[ \tau \text{ unbounded , if } \mu \bar{\tau} \leq -1 \quad (2.3.a) \]

\[ \tau \leq \tilde{\tau} \Psi(\mu \bar{\tau}) \text{ , if } \mu \bar{\tau} > -1 \quad (2.3.b) \]

where \( \bar{\tau} := (2L_0L_2)^{-1/2} \) and

\[ \Psi(s) := \begin{cases} \ln(1+s)/s & s \neq 0 \\ 1 & s = 0 \end{cases} \quad (2.3.c) \]
Proof. One applies Theorem 3 from [10]. Let $|\cdot|$ denote the standard $C^0$-norm:

$$|u| := \max_{t \in [0, T]} \|u(t)\|,$$

where $\|\cdot\|$ is induced by $\langle \cdot, \cdot \rangle$. In the above introduced notation, Theorem 3 essentially requires that

$$|\Delta y^0| \leq \alpha \quad (2.4.a)$$

$$|F_y(y^0)^{-1}(F_y(u) - F_y(v))| \leq \omega |u - v| \quad (2.4.b)$$

$$\alpha \omega \leq \frac{1}{2} \quad (2.4.c)$$

The rest of the assumptions holds for $D$ sufficiently large. The task is now to estimate $\alpha, \omega$ and to apply (2.4.c). With $y^0(t) \equiv y_0$, the first Newton correction satisfies - compare (2.1):

$$\Delta y^0(r) = \int_{t=0}^r \exp(A(r - t))f(y_0)dt$$

Hence

$$\|\Delta y^0(r)\| \leq \int_{t=0}^r \|\exp(As)f(y_0)\|ds \leq L \int_{t=0}^r \exp(\mu s)ds = L_0 \tau \varphi(\mu \tau) =: \alpha(\tau)$$

with $\varphi$ as introduced in (1.7.b).

In order to estimate $\omega(\tau)$, one introduces the operator norm in (2.4.b) by

$$z := F_y(y^0)^{-1}(F_y(v + w) - F_y(v))u$$

$$|z| \leq \omega \cdot |u| \cdot |w|$$

Once more by variation of constants, one obtains

$$\|z(\tau)\| \leq \int_{t=0}^\tau \|\exp(A(\tau - t))[f_y(v + w) - f_y(v)]u\|dt ,$$

which, similar as above, yields

$$\|z(\tau)\| \leq L_2 \cdot \tau \cdot \varphi(\mu \tau) \cdot |u| \cdot |w|$$

Hence, a natural definition is

$$\omega(\tau) := L_2 \tau \varphi(\mu \tau) .$$

Insertion into the Kantorovitch condition produces

$$(\tau \varphi(\mu \tau))^2 \leq (2L_0L_2)^{-1} =: \tilde{r}^2 \quad (2.5)$$
or, equivalently,

\[ r \varphi(\mu r) \leq \tau. \quad (2.5') \]

Then (2.3) is an immediate consequence.

Theorem 1 gives a clear description of the local continuation property of solutions of (1.1) in terms of \( \mu \) and \( \tau \) - but under the assumption of a given exact Jacobian \( f_y(y_0) \). In many applications, however, an approximation error

\[ \delta A := A - f_y(y_0) \]

must be taken into account.

**Theorem 2** Notation and assumptions as in Theorem 1, but with

\[ \|\delta A\| \leq \delta_0, \quad \delta_0 \geq 0. \quad (2.7) \]

Then the results (2.3) hold with \( \tau \) replaced by

\[ \hat{\tau} := \frac{\tau}{1 + \delta_0 \tau} \]

**Proof.** Once more, Theorem 3 of [10] is applied with \( F_y(y_0) \) replaced by \( M_F(y_0) \), which means replacing \( f_y(y_0) \) by \( A \neq f_y(y_0) \). With \( \mu \) now the logarithmic norm of the Jacobian approximation \( A \), the estimates for \( \alpha(r), \omega(\tau) \) carry over. In addition, the assumptions (2.4) must be extended by

\[ \|M_F(y_0)^{-1}(F_y(y_0) - M_F(y_0))\| \leq \delta_0 < 1 \quad (2.9.a) \]

Upon defining

\[ z := M_F(y_0)^{-1}(F_y(y_0) - M_F(y_0))u, \]

a similar estimate as in the proof of Theorem 1 leads to

\[ ||z(r)|| \leq \int_0^r \|\exp(A(r - t) \cdot \delta A \cdot u\|dt \leq \delta_0 r \varphi(\mu r)|u| \]

Hence, one obtains the condition

\[ \delta_0 := \delta_0 r \varphi(\mu r) < 1 \quad (2.9.b) \]

Insertion into the modified Kantorovitch condition

\[ \frac{\alpha \omega}{(1 - \delta_0)^2} \leq \frac{1}{2} \quad (2.9.c) \]

then yields

\[ r \varphi(\mu r) \leq \tau/(1 + \delta_0 \tau) =: \hat{\tau} \]

Note that condition (2.9.b) is automatically satisfied.

The condition

\[ \mu \hat{\tau} \leq -1 \Leftrightarrow \mu + \hat{\tau}^{-1} \leq 0 \quad (2.10) \]

may be regarded as a characterization of **contractive ODE systems.**
3. Stepsize Bounds for Implicit Discretization Methods

The quantities $\mu, \tau, \delta_0$ defined in the preceding section are now applied to determine stepsizes bounds for implicit one-step discretization methods. The formalism can be sufficiently exemplified by the implicit Euler discretization. In each step of this discretization, one must solve the algebraic system

$$F(y, h) := y - y_0 - hf(y) = 0,$$  \hspace{1cm} (3.1)

which now represents a homotopy in $\mathbb{R}^n$ with embedding in terms of the stepsize $h$ - say $h \geq 0$. The Newton-like iteration for solving this system is

$$(I - hA)\Delta y_i = -(y_i - y_0 + hf(y_i)),$$  \hspace{1cm} (3.2)

where $\delta A := A - f_y(y_0) \neq 0$ will be assumed.

**Theorem 3** Assumptions and notation as in Theorem 1 and Theorem 2 above. Then the Newton-like iteration (3.2) for the implicit Euler discretization converges to a unique solution for all stepsizes.

$h$ unbounded, if $\mu^\tau \leq -1$  \hspace{1cm} (3.3.a)

$h \leq \tau \Psi_{\text{IE}}(\mu^\tau)$, if $\mu^\tau > -1$  \hspace{1cm} (3.3.b)

where

$$\Psi_{\text{IE}}(s) := (1 + s)^{-1}$$  \hspace{1cm} (3.3.c)

**Proof.** Once more, Theorem 3 of [10] is applied, here to the homotopy (3.1). The Jacobian approximation $A \approx f_y(y_0)$ leads to the approximation

$$I - hA =: M_F(y_0) \approx F_y(y_0),$$

which is used in the definition of the affine invariant Lipschitz constant

$$\|M_F(y_0)^{-1}(F_y(u) - F_y(v))\| \leq \omega(h)\|u - v\|,$$  \hspace{1cm} (3.4.a)

the first correction bound

$$\|\Delta y_0\| = \|M_F(y_0)^{-1}F(y_0)|\| \leq \alpha(h),$$  \hspace{1cm} (3.4.b)

and the approximation measure

$$\|M_F(y_0)^{-1}(M_F(y_0) - F_y(y_0))\| \leq \bar{\delta}_0(h) < 1$$  \hspace{1cm} (3.4.c)
With these definitions, the modified Kantorovitch condition reads

\[
\frac{\alpha \omega}{(1 - \delta_0)^2} \leq \frac{1}{2} .
\]  

\hspace{1cm} (3.4.d)

Upon using similar techniques as in the proofs of Theorem 1 and Theorem 2 above, one comes up with the estimates:

\[
\alpha(h) := \frac{hL_0}{1 - \mu h} \\
\omega(h) := \frac{hL_2}{1 - \mu h} \\
\delta_0(h) := \frac{h\delta_0}{1 - \mu h}
\]

where \(L_0, L_2, \delta_0\) are defined as in section 2. Insertion into (3.4.d) yields, for \(\mu h < 1\):

\[
\frac{h}{1 - \mu h} \leq \bar{\tau},
\]

or, equivalently,

\[
h \leq \bar{\tau}/(1 + \mu \bar{\tau}) .
\]  

This is just (3.3). Finally, note that for \(\mu > 0\)

\[
\mu h \leq \mu \bar{\tau}/(1 + \mu \bar{\tau}) < 1 ,
\]

which assures the above requirement.

In Figure 1, the comparable functions \(\Psi_{\text{IE}}(s)\) from Theorem 3 and \(\Psi(s)\) from Theorem 1 and 2 are represented.

Figure 1 Comparison of functions \(\Psi\) (analytic case) and \(\Psi_{\text{IE}}\) (implicit Euler discretization).
Common features are

\[ a) \quad \psi(0) = \psi_{IE}(0) = 1 , \]
\[ b) \quad \psi \text{ and } \psi_{IE} \text{ have a pole at } s = -1 , \quad (3.7) \]
\[ c) \quad \psi \text{ and } \psi_{IE} \text{ are monotonically decreasing.} \]

Differences are:

\[ a) \quad \psi'(0) = -\frac{1}{2} , \quad \psi'_{IE}(0) = -1 \]
\[ b) \quad s > 0 : \quad \psi_{IE}(s) < \psi(s) \quad (3.8) \]
\[ c) \quad -1 < s < 0 : \quad \psi_{IE}(s) > \psi(s) . \]

Note that (3.8.c) indicates the possible occurrence of spurious solutions from the implicit Euler discretization.

The generalization of the above results to implicit one-step methods is immediate. One just observes that

\[ R_{IE}(z) = (1 - z)^{-1} , \quad z \in \mathbb{C} \]

is the stability function of the implicit Euler method. Let \( R(z) \) be the stability function of a general one-step method. Then \( R(hA)y_0 \) characterizes the discrete solution obtained for \( f = Ay \) - and therefore represents the solution of the associated linear system. Hence, in the nonlinear case, the convergence of a typical Newton-like iteration can be discussed in terms of assumptions just like (3.4) with

\[ M_R(y_0)^{-1} \rightarrow R(hA) . \quad (3.9) \]

Following [15], one may introduce

\[ \varphi_R(s) := \sup_{R(z) \leq s} |R(z)| . \quad (3.10) \]

Then the Kantorovich condition leads to

\[ h\varphi_R(\mu h) \leq \delta , \quad (3.11) \]

which thus implicitly defines some function \( \Psi_R(s) \) so that (3.11) is equivalent to

\[ h \leq \delta \Psi_R(\mu \delta) . \quad (3.11') \]

Of course, the function \( \Psi_R \) associated with the discretization method should model the function \( \psi \) associated with the analytic ODE-IVP - compare (3.7) for the special case \( \Psi_R = \psi_{IE} \). First, the condition

\[ \Psi_R(0) = \Psi(0) = 1 , \]
requires that
\[ \varphi_R(0) = 1 . \] (3.12.a)

This property will hold, whenever
\[ |R(iv)| \leq |R(0)| = 1 , v \in \mathbb{R}^1 , \]
which is certainly true for A-stable discretizations. Next, \( h \) should be permitted to be unbounded for contractive ODE systems - which means that \( \Psi_R \) has a pole somewhere on the negative real axis. In this case, condition (3.11) requires that
\[ \varphi_R(\mu h) \leq \frac{t}{h} \text{ for } h \to \infty , \] (3.12.b)
which implies the linear stability condition
\[ R(-\infty) = 0 . \] (3.13)

Thus, the requirements of L-stability naturally come out of the Kantorovitch conditions for the analytic and the discretized ODE-IVP.

Remark 1. The author is aware of the fact that the above characterization exhibits similarities with work of Frank/Schneid/Überhuber [13]. However, in order to study the local order of stiff integrators, the above stepsize bounds should certainly be included into their theory.

Remark 2. For multistep methods, a natural conjecture is that the Kantorovitch condition leads one to generalize condition (3.13) to
\[ \zeta_j(-\infty) = 0 \text{ for } j = 1, \ldots, p , \] (3.14)
where \( \zeta_j(z) \) are the characteristic roots of the equation \( (\zeta \in \mathbb{C}, z \in \mathbb{C}) \)
\[ \rho(z) - z \sigma(z) = 0 , \]
and \( (\rho, \sigma) \) are the standard polynomials determining the multistep method (of order \( p \)) under consideration. It is well-known that condition (3.14) directly leads to the BDF discretization, if only the minimal set of free parameters is used - this is Gear's method [14]. However, the requirement of A-stability confines one to \( p \leq 2 \), if (3.12.a) is wanted.

Remark 3. By construction, \( R(z) \) is an approximation of \( \exp(z) \). However, by comparison of (2.5) and (3.11), one observes that, in the present context, \( R(z) \) is serving as an approximation of \( \varphi(z) = (\exp(z) - 1)/z \). Therefore, a simultaneous approximation should satisfy
\[ (R(z) - i)/z = R(z) . \] (3.15)
This property only holds for the implicit or semi-implicit Euler discretization.
4. Computational Estimation of Stepsize Bounds in Stiff Extrapolation Integrators

At present, the most efficient integrators of extrapolation type are based on

a) the semi-implicit midpoint rule [2] - code METAN1,

b) the semi-implicit Euler discretization [7] - code EULSIM,

c) the modified semi-implicit Euler discretization for implicit and differen­tial-algebraic systems [9,11] - code LIMEX.

These discretizations are repeated successively over a basic step \([0, H]\) with internal stepsizes

\[ h_i := H/n_i , \quad n_i \in \mathcal{F} , \]

where \( \mathcal{F} \) is an integer sequence defined in the codes [7]. Just due to this repetition structure, easy and cheap estimates of \( \mu, \tau, \delta_0 \) are available within the above codes.

4.1 Estimation of Jacobian Logarithmic Norm

Let \( |\mu| \) denote a computational estimate of the Jacobian logarithmic norm \( \mu \). Both discretizations (a) and (b) start with a semi-implicit Euler step by computing

\[ \Delta y_0(h) := h(I - hA)^{-1}f(y_0) \]  

(4.1)

where \( A \approx f_y(y_0) \). Thus, the quantities

\[ d_i := ||\Delta y_0(h_i)|| \]  

(4.2)

are available, at least for \( i = 1, 2 \). Let the norm be induced by an inner product, e.g. the Euclidean norm or, in the method of lines, an appropriate discrete Sobolev-norm. Then, in the above introduced notation, one obtains

\[ d_i \leq h_i L_0 / (1 - \mu h_i) \]  

(4.3.a)

under the condition

\[ \mu h_i < 1 . \]  

(4.3.b)

Now, define the computationally available quantities

\[ \kappa_i := d_i / d_{i-1} , \text{ if } d_{i-1} \neq 0 \]

\[ \bar{\kappa}_i := h_i / h_{i-1} = n_{i-1} / n_i . \]
By definition, one has \( \kappa_i < 1 \). Moreover, in order to make stepsize refinement a reasonable device, the condition \( \kappa_i < 1 \) is natural. Under this assumption, (4.3) leads to the desired estimate

\[
\mu h_i = \frac{\kappa_i - \kappa_i}{1 - \kappa_i} =: [\mu_i] h_i,
\]

which implies (from \( \kappa_i > 0 \))

\[
[\mu_i] h_{i-1} < 1.
\]

The generalization to implicit ODE's

\[
B(y)y' = f(y)
\]

is immediate, if \( B \) is a nonsingular \((n,n)\)-matrix (\( \text{index} = 0 \)). The first step of discretization (c) computes (with \( B_0 := B(y_0) \)):

\[
\Delta y_0(h) := h(B_0 - hA)^{-1} f(y_0)
\]

Under the assumption

\[
(\Delta y_0, B_0 \Delta y_0) \geq b_0 (\Delta y_0, \Delta y_0)
\]

the appropriate extension of (4.3) can be derived:

\[
a) \quad d_i \leq h_i L_0 / (b_0 - \mu h_i), \\
b) \quad \mu h_i < b_0
\]

Hence, under the assumption

\[
b_0 > 0,
\]

the results (4.3) and (4.4) hold with the mere replacement:

\[
(L_0, \mu) \rightarrow (L_0/b_0, \mu/b_0).
\]

Remark. \( B \) positive-definite is often satisfied in problems of type (4.5), which arise from finite element method of lines treatment of parabolic PDE's.

For singular matrix \( B \) a differential-algebraic system occurs (\( \text{index} > 0 \)). Let \( B^+ \) denote the Moore-Penrose pseudoinverse of \( B \). Then the algebraic part of (4.5) can be written in the form [11]:

\[
(I - B(y)B^+(y))f(y) = 0
\]

Let the initial values be consistent, which means that (4.11) holds for \( y_0 \).

In addition, assume \( \text{index} = 1 \). Then

\[
\lim_{h \rightarrow 0} (\frac{1}{h} \Delta y_0(h)) = B_0^+ f(y_0)
\]
Proof: For index \( \eta = 1 \), Theorem 5 of [22] states that, in the present notation

\[
\lim_{h \to 0} (B_0 - hA)^{-1} B_0 = B_0^+ B_0 .
\]

If \( y_0 \) is consistent, then

\[
\frac{1}{h} \Delta y_0 = (B_0 - hA)^{-1} f(y_0) = (B_0 - hA)^{-1} B_0 B^+ f(y_0)
\]

\[
\rightarrow B_0^+ B_0 B^+ f(y_0) = B_0^+ f(y_0)
\]

As a consequence, \( (4.7) \) for \( h \to 0 \), approaches

\[
\langle B_0^+ f(y_0), B_0 B_0^+ f(y_0) \rangle \geq b_0 \langle B_0^+ f(y_0), B_0^+ f(y_0) \rangle . \quad (4.7')
\]

Hence, condition \( (4.9) \) is a reasonable assumption also for singular \( B_0 \) as long as \( \text{index} = 1 \) is guaranteed.

4.2 Estimation of Jacobian Perturbation and Jacobian Lipschitz Constant

This section deals with the question of how to obtain cheap computational estimates of the above introduced theoretical quantities \( \delta_0 \) and \( \tau \). Since \( \tau \) contains second order information, at least two Newton iterations will be required in the implicit Euler discretization. Let \( \Delta y^0 := \Delta y_0 \) from \( (4.1) \) and \( y^1 := y_0 + \Delta y_0 \). Then the second Newton correction reads

\[
\Delta y^1 := (I - hA)^{-1} (h f(y^1) - \Delta y^0) . \quad (4.12)
\]

Straightforward Taylor expansion yields

\[
\Delta y^1 = h(I - hA)^{-1} w
\]

\[
w := -\delta A \Delta y_0 + \frac{1}{2} \int_{s=0}^{1} f_{yy}(y_0 + s \cdot \Delta y_0) \Delta y_0 \Delta y_0 ds . \quad (4.13)
\]

Denote the contraction factor by

\[
\Theta := \|\Delta y^1\| / \|\Delta y^0\| . \quad (4.14)
\]

Then standard estimation as in the preceding section 3 leads to

\[
\Theta \leq \delta_0 \cdot z + (z/2\tau)^2
\]

\[
z := h/(1 - \mu h) \quad (4.15)
\]

14
Combining (4.15) with (3.6), one obtains the maximum contraction factor

\[ \Theta_{\text{max}} = \frac{1}{4} \left( \frac{1 + 2\delta_0\tilde{\tau}}{1 - \delta_0\tilde{\tau}} \right)^2, \]

which nicely reflects the limiting cases \( \Theta_{\text{max}} = \frac{1}{4} \) (simplified Newton method) and \( \Theta_{\text{max}} = 1 \) (fixed point iteration). Upon requiring \( \Theta \leq \Theta_{\text{max}} \) for some default value \( \Theta_{\text{max}} \), one comes up with the (possible) restriction

\[
\hat{\tau} = \frac{4\hat{\delta}\tau / (\alpha + \sqrt{4\hat{\delta} + \alpha^2})}{4\hat{\delta}\tau}
\]

as a sufficient contractivity condition for the implicit Euler method.

In the semi-implicit discretizations (a) to (c), the first question is how to compute \( \Delta y^1 \) in a cheap way. For the semi-implicit midpoint discretization (METAN1) one conveniently uses the identity

\[ \Delta y^1 = (\Delta y^0 - \Delta y^0) / 2 \]

as already observed in [7], formula (3.28). For the semi-implicit Euler discretization (EULSIM), the actual performance of (4.12) needs an extra forward/backward substitution, when \( n_t > 1 \). For the modified semi-implicit Euler discretization (LIMEX, index = 0, 1), the representation (4.12) must be modified to

\[
\Delta y^0 := h(B_0 - hA)^{-1}f(y_0) \\
y^1 := y_0 + \Delta y^0, \quad (4.12') \\
\Delta y^1 := (B_0 - hA)^{-1}(hf(y^1) - B(y^1)\Delta y^0).
\]

Once more, an extra substitution is needed to obtain the second order information.

Assume now that \( [\mu] \) is already computed. Then the desired estimates \([\hat{\tau}], [\delta_0]\) could, in principle, be obtained from the model assumption

\[ \Theta = \delta_0 \cdot z + (z/2\hat{\tau})^2 \]  

(4.17)

and values \( \Theta(h_f), \Theta(h_{f+1}) \). However, (4.17) requires positivity constraints on the coefficients. Therefore, \( \Theta \) is replaced by the sign-dependent quantity

\[ \hat{\Theta} := \frac{\langle \Delta y^0, \Delta y^1 \rangle}{\langle \Delta y^0, \Delta y^0 \rangle} \]

(4.18)

satisfying

\[ |\hat{\Theta}| \leq \Theta. \]
Then, under the model assumption

\[ \Theta(z) = \beta \cdot z + \gamma \cdot z^2 \]  

(4.19.a)

with \( \beta, \gamma \) of either sign, one may obtain the estimates

\[ \| \delta_0 \| := |\beta|, [\tau^{-1}] := 2\sqrt{|\gamma|} \]  

(4.19.b)

from values \( \Theta(h_j), \Theta(h_{j+1}). \)

If \( \mu h \ll -1 \), then (4.17) degenerates essentially to

\[ \Theta \lessapprox \delta_0 /[|\mu| + (2|\mu| |\tau|)^{-2}] \]  

(4.17')

so that the \( h \)-dependence of the right-hand side has practically vanished.

In order to gain distinguishable information from the linear and quadratic term, the estimates (4.19) are only trusted when

\[ \frac{|\mu h|}{|1 - \mu h|} \leq \frac{1}{3} \]  

(4.20)

or equivalently, when

\[ -1 \leq \mu h \leq \frac{1}{3} \]  

(4.20')

Of course, the test (4.20') is made replacing \( \mu \) by [\( \mu \)].

Finally, recall from [7] that second order information can also be used to decide when to actually regenerate the Jacobian matrix approximation \( A \).
5. Numerical Experiments

The devices described in section 4 above were implemented in the extrapolation codes METAN1, EULSIM, LIMEX. Besides the standard order and stepsize control due to [6], different variants using different additional order and stepsize constraints were realized. These variants are

OLD: old empirical devices (former standard options).

MY: stepsize prediction subject to $\mu H \leq c_1 \cdot n_1$ (with $c_1 = 1/2$ in EULSIM and LIMEX, $c_1 = 1/3$ in METAN1); stepsize reduction if $\mu H > 0.9n_1$.

MYJ: MY-version + $\Theta$-computation (once per basic step); Jacobian regeneration only, if $\Theta > \Theta_{\min}$; no order increase in the next basic step, if $\Theta > 1/8$; stepsize reduction, if $\Theta > 1/4$.

MYT: MYJ-version with $\Theta \to \bar{\Theta}$ (twice per basic step); stepsize restriction (4.16) with $\bar{\Theta} = 1/8$.

Jacobian numerical differencing is used throughout. All computations have been done on the SIEMENS 7.865 of the Konrad-Zuse-Center, Berlin. Generally speaking, it is not at all clear whether the semi-implicit methods require stepsize restrictions such as (4.16). However, the analytic problem has similar restrictions. As a consequence, one will expect comparable restrictions to show up in the asymptotic expansions of the semi-implicit discretizations. Investigations in this direction should be made - they are, however, certainly beyond the scope of this paper.

STIFF DETEST [12]. In this well-known test set, the old and new versions of EULSIM and METAN1 behaved rather similarly - showing improvements in the 5% level. Even the Jacobian evaluation savings did not pay off significantly in terms of computing time (of course, in terms of $f$-evaluations in the linear classes $A, B$). Detailed comparison numbers are not worth giving. Rather, more challenging examples will be included below.
Example 1. Van der Pol's equation. The example including initial values was taken from [16] for moderate user prescribed tolerances TOL and parameters ε.

Table 1 Comparison of number NFEV of f-evaluations in example 1 for different variants of the extrapolation code EULSIM.

<table>
<thead>
<tr>
<th>ε \ TOL</th>
<th>10^{-2}</th>
<th>10^{-4}</th>
<th>10^{-6}</th>
</tr>
</thead>
<tbody>
<tr>
<td>10^{-2}</td>
<td>OLD</td>
<td>166</td>
<td>469</td>
</tr>
<tr>
<td></td>
<td>MY</td>
<td>153</td>
<td>407</td>
</tr>
<tr>
<td></td>
<td>MYJ</td>
<td>160</td>
<td>398</td>
</tr>
<tr>
<td></td>
<td>MYT</td>
<td>156</td>
<td>392</td>
</tr>
<tr>
<td>10^{-4}</td>
<td>OLD</td>
<td>286</td>
<td>780</td>
</tr>
<tr>
<td></td>
<td>MY</td>
<td>247</td>
<td>713</td>
</tr>
<tr>
<td></td>
<td>MYJ</td>
<td>274</td>
<td>692</td>
</tr>
<tr>
<td></td>
<td>MYT</td>
<td>258</td>
<td>696</td>
</tr>
<tr>
<td>10^{-6}</td>
<td>OLD</td>
<td>411</td>
<td>1209</td>
</tr>
<tr>
<td></td>
<td>MY</td>
<td>331</td>
<td>932</td>
</tr>
<tr>
<td></td>
<td>MYJ</td>
<td>373</td>
<td>960</td>
</tr>
<tr>
<td></td>
<td>MYT</td>
<td>352</td>
<td>962</td>
</tr>
</tbody>
</table>

In addition to the above test results, experiments with the stepsize sequence suggested in [16] for the semi-implicit Euler method with extrapolation have been performed: they demonstrated a drastic slowing down for the test values of (ε, TOL) herein.

The relative gain from version OLD to version MY, MYJ or MYT was similar for the code METAN1. However, in this example, EULSIM clearly superceded METAN1 up to TOL=10^{-6}.
Example 2. Kidney model problem. A documentation and illustration of this test problem can be found in [2]. The value \( y(0) = 0.99 \) is selected for comparison. The example is known to be extremely sensitive to Jacobian approximation errors.

Table 2 Comparison of computing times and \( f \)-evaluations for variants of E(ULSIM) and M(ETAN1) in example 2 with numerical Jacobian used in the codes.

<table>
<thead>
<tr>
<th>TOL</th>
<th>TIME</th>
<th>NFEV</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 10^{-4} )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E-OLD</td>
<td>0.77</td>
<td>634</td>
</tr>
<tr>
<td>E-MY</td>
<td>0.64</td>
<td>487</td>
</tr>
<tr>
<td>E-MYJ</td>
<td>0.66</td>
<td>487</td>
</tr>
<tr>
<td>E-MYT</td>
<td>0.75</td>
<td>491</td>
</tr>
<tr>
<td>M-OLD</td>
<td>0.84</td>
<td>1002</td>
</tr>
<tr>
<td>M-MY</td>
<td>0.59</td>
<td>706</td>
</tr>
<tr>
<td>M-MYJ</td>
<td>0.62</td>
<td>672</td>
</tr>
<tr>
<td>M-MYT</td>
<td>0.52</td>
<td>593</td>
</tr>
<tr>
<td>( 10^{-6} )</td>
<td>1.36</td>
<td>1191</td>
</tr>
<tr>
<td>E-OLD</td>
<td>1.14</td>
<td>917</td>
</tr>
<tr>
<td>E-MY</td>
<td>1.16</td>
<td>917</td>
</tr>
<tr>
<td>E-MYT</td>
<td>1.11</td>
<td>780</td>
</tr>
<tr>
<td>M-OLD</td>
<td>1.20</td>
<td>1493</td>
</tr>
<tr>
<td>M-MY</td>
<td>1.06</td>
<td>1318</td>
</tr>
<tr>
<td>M-MYJ</td>
<td>1.11</td>
<td>1320</td>
</tr>
<tr>
<td>M-MYT</td>
<td>0.86</td>
<td>996</td>
</tr>
</tbody>
</table>

Obviously, significant reductions are obtained. In terms of \( f \)-evaluations, EULSIM-MYT is most economic. The cost of evaluating \( f \), however, is so small that METAN1-MYT is faster.
Example 3. RNA-polymerization. A documentation and illustration of this large scale problem is given in [3] in the context of the chemical kinetics package LARKIN. At that time, LARKIN used an appropriate adaptation of METAN1 for stiff integration. A more recent revision due to [20] allows both for EULSIM and METAN1 versions. Recall that LARKIN generates the sparse analytic Jacobian fairly cheaply, so that savings from keeping the Jacobian are not expected. This feature and the comparisons in Tables 1 and 2 seem to indicate that testing the EULSIM-MY version is the interesting part.

Table 3 Comparison for example 3, EULSIM adaptation within LARKIN (TOL=1.D-3).

<table>
<thead>
<tr>
<th></th>
<th>NFEV</th>
<th>NDEC</th>
<th>NSUBST</th>
</tr>
</thead>
<tbody>
<tr>
<td>E-OLD</td>
<td>995</td>
<td>425</td>
<td>1345</td>
</tr>
<tr>
<td>E-MY</td>
<td>784</td>
<td>380</td>
<td>1083</td>
</tr>
</tbody>
</table>

The gain in this example is essentially obtained by indirect restriction of order increase in the extrapolation code, since the constraint $\mu H \leq c_1 \cdot n_1$ is activated throughout most of the integration. The positive $\mu$ herein nicely reflects the "polymerization wave" [3]. The total computing time in this large scale example follows the behavior of NSUBST, the number of forward/backward substitutions.
Example 4. Hydrogen-Oxygen Auto-Ignition. This example comprises a physical and chemical model due to [19]. The model describes the auto-ignition process of hydrogen-oxygen mixtures in space and time. The example presented here is a 1D-system of reaction-diffusion-convection type. Method of lines preprocessing (with strongly non-uniform grid) leads to a differential-algebraic system of index = 1, which is treated by the extrapolation code LIMEX. For testing purposes, the MY-version of LIMEX was compared with the old version already implemented within the flame codes of [19].

Table 4 Comparison of computing times for example 4, LIMEX adaptation within a flame code due to [19], TOL=1.D-3.

<table>
<thead>
<tr>
<th></th>
<th>(t_{fin}=2.0D-2)</th>
<th>(t_{fin}=1.0D3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIMEX-OLD</td>
<td>952 sec</td>
<td>2233 sec</td>
</tr>
<tr>
<td>LIMEX-MY</td>
<td>433 sec</td>
<td>1525 sec</td>
</tr>
</tbody>
</table>

The run of interest for the physical chemist is the one up to final time \(t_{fin}=2.0D-2\) – which is shortly after the auto-ignition point. At \(t_{fin}=1.0D3\), the stationary state of the system has been reached. It may be worth noting that investigations of this kind help to quantitatively understand the phenomenon of “engine knock”.

21
Conclusion

The above theoretical characterization of stiff ODE initial value problems leads to additional stepsize bounds in stiff integrators – which can be computed especially cheaply in extrapolation codes. The numerical experiments demonstrate that (a) stepsize restrictions due to the Jacobian logarithmic norm $\mu$ are worth implementing and lead to a significant speed-up of computations, in particular in combustion problems, (b) stepsize restrictions due to Jacobian perturbations and Jacobian Lipschitz constant are essentially automatically respected by stiff integrators – thus nicely demonstrating the effective description in terms of the underlying theory.

Acknowledgments. The author wishes to thank U. Pöhle for extensive computational assistance and U. Nowak and U. Maas for helpful discussions.
References


