The version of the following full text has not yet been defined or was untraceable and may differ from the publisher's version.

For additional information about this publication click this link.
http://hdl.handle.net/2066/18713

Please be advised that this information was generated on 2019-08-17 and may be subject to change.
Stability and error estimates for the \( \theta \)-method for strongly monotone and infinitely stiff evolution equations. *

O.AXELSSON † S.V.GOLOLOBOV ‡

March 16, 1999

Abstract

For evolution equations with a monotone operator \( F(t, u(t)) \) we derive unconditional stability and discretization error estimates valid for all times. For the \( \theta \)-method, with \( \theta = 1 - \frac{1}{2 + \nu} \), \( 0 < \nu \leq 1 \), \( \zeta > 0 \), we prove an error estimate \( O(\tau^{\frac{1}{2}}) \), \( \tau \rightarrow 0 \), if \( \nu = \frac{1}{2} \), where \( \tau \) is the maximal step for an arbitrary choice of sequence of steps and with no assumptions about the existence of the Jacobian as well as other derivatives of the operator \( F(\cdot, \cdot) \), and an optimal estimate \( O(\tau^2) \) under some additional relation between neighboring steps. The first result is an improvement over the implicit midpoint method \( \theta = \frac{1}{2} \), for which an order reduction to \( O(\tau) \) may occur for infinitely stiff problems. Numerical tests illustrate the results.

Mathematics Subject Classification (1991): 65J05, 65J15, 65L20, 65L70

1 INTRODUCTION

The present paper is based mainly on the paper [3] the results of which were not widely published. It also includes another treatment of the optimal order estimates presented in Section 3.2 and some numerical examples, that confirms the results obtained.

Consider the evolution equation

\[
\frac{du}{dt} + F(t, u(t)) = 0, \quad t > 0, \quad u(0) = u_0 \in V,
\]

where \( V \) is a reflexive Banach space, \( u_t = \frac{du}{dt} \) and \( F(t, \cdot) : V \rightarrow V' \). Here \( V' \) is the space which is dual with respect to the inner product \( (\cdot, \cdot) \) in a Hilbert space \( H \), with norm \( \|v\| = (v, v)^{\frac{1}{2}} \).

*This work has been supported by the NWO program, doss.nr. 047-003-017, and by the Russian Foundation for Basic Research (98-01-00709).

†Department of Mathematics, Catholic University of Nijmegen, Toernooiveld 1, 6525 ED Nijmegen, The Netherlands. E-mail: axelsson@sci.kun.nl.

‡Institute of Computational Mathematics and Mathematical Geophysics, Siberian Branch of Russian Academy of Sciences, Lavrentiev av. 6, 630090 Novosibirsk, Russia.
We shall assume that $F(\cdot, \cdot)$ is a monotone operator, i.e.

$$(F(t, u(t)) - F(t, v(t)), u(t) - v(t)) \geq \rho(t)\|u(t) - v(t)\|^2, \quad \forall u(t), v(t) \in V,$$  \hspace{1cm} (2)

where $\rho(\cdot) : (0, \infty) \to \mathbb{R}^+, \text{ i.e. } \rho(t) \geq 0, \quad t > 0.$

A typical example is the parabolic evolution equation

$$u_t = \nabla \cdot (a(t, x, \nabla u) \nabla u) + g(t, u), \quad (t, x) \in (0, \infty) \times \Omega, \quad \Omega \subset \mathbb{R}^d,$$  \hspace{1cm} (3)

with boundary conditions, say $u = 0$ on $\partial \Omega$. Here $V = \left[ L^2(\Omega) \right]^d$ (a Cauchy product of the Sobolev space $H^1(\Omega)$, $H = L^2(\Omega)$ and, under certain conditions on $a(\cdot, \cdot, \cdot)$ and $g(\cdot, \cdot)$ this is a parabolic problem, i.e. fulfills (2) with $\rho(t) > 0$.

Other important examples are conservative (hyperbolic) problems for which (2) is satisfied with $\rho(t) \geq 0$. In the present paper we restrict the analysis to the strongly monotone case, $\rho(t) > \rho_0 > 0$.

Classical techniques for the derivation of discretization error estimates for (3) uses a semidiscrete method for the discretization in space, namely the variational form

$$(u_t, v) + (F(t, u), v) = 0, \quad \forall v \in V_h \subset V,$$

where $V_h$ is a finite element space depending on a mesh parameter $h$.

This semidiscrete method ("longitudinal method of lines") results in a system of ordinary differential equations (ODE) which is stiff, i.e. components of the solution exist, which decay (exponentially) with largely different rates.

The system of ODE can be solved by many methods for stiff ODE’s. The difficulty is in proving error estimates for the total error of the form $C_1 h^p + C_2 \tau^q, \quad p, q > 0$, where $\tau$ is the step. Here $C_1, C_2$ should be independent of $\tau, h$. Since the dimension of the ODE depends on $h$, classical error estimates used in numerical analysis of ODE, cannot be applied. Furthermore, they provide usually only a bound growing with time $t$ (sometimes even growing exponentially - see below). We want to derive error estimates which are valid (i.e. bounded) for all $t$.

We find it then convenient to consider a "transversal method of lines", i.e. first discretize the evolutionary problem (1) with respect to time. We must then be able to handle unbounded operators. A convenient time integration method turns out to be the implicit ("one-leg") form of the $\theta$-method with $1 - \frac{\nu(\zeta - \theta)}{2 + \zeta} \leq \theta \leq 1, \quad \zeta > 0$ for some $\nu, \quad 0 \leq \nu \leq 1$. Error estimates $O(\tau^{\frac{\zeta - \nu}{2 + \zeta}})$ valid for all $t$ can now be derived in the strongly monotone case, where $\rho(t) \geq \rho_0 > 0$. Without further assumptions the optimal order we prove is $O(\tau^{\frac{1}{2}})$ for $\nu = \frac{1}{3}$ and $\theta$ equal to the upper bound. With some additional assumptions we prove also the optimal order, $O(\tau^2)$.

To illustrate the problem with proving error estimates for time-stepping methods, we consider first the Euler (forward) method,

$$v(t + \tau) = v(t) - \tau F(t, v(t)), \quad t = 0, \tau, 2\tau, \ldots,$$  \hspace{1cm} (4)
where $v$ is the corresponding approximation to $u$. (It is only for notational simplicity that we let the step $\tau$ be constant.)

Let $e(t) = u(t) - v(t)$ be the error function. Classical error estimates use the two-sided Lipschitz constant,

$$L = \sup_{t > 0} \frac{\|F(t, u) - F(t, v)\|}{\|u - v\|}, \quad u, v \in V_0 \subset V,$$

where $V_0$ contains all functions in a sufficiently large tube about the solution $u(t)$. In the analysis of the Euler forward method we have to assume that $F(\cdot, \cdot)$ is two-sided Lipschitz bounded, i.e. that $L < \infty$, but for the implicit methods to be considered later, we need only a one-sided bound such as (2). From (1) it follows

$$u(t + \tau) = u(t) - \int_0^1 F(t + \tau s, u(t + \tau s)) ds$$

and from (4), (5) we get

$$e(t + \tau) = e(t) - \tau \left( F(t, u(t)) - F(t, v(t)) \right) + \tau R(t, u(t)),$$

where

$$R(t, u(t)) = \int_0^1 (F(t, u(t)) - F(t + \tau s, u(t + \tau s))) ds$$

$$= \int_0^1 (u(t + \tau s) - u(t)) ds$$

is the (normalized) local truncation error. Note that

$$\sup_{t > 0} \|R(t, u(t))\| = \tau \sup_{t > 0} \int_0^1 ds \int_0^\tau \|u^{(2)}(t + \sigma \tau)\| d\sigma \leq \frac{1}{2} \tau D_2,$$

where we use the notation

$$D_k = \sup_{t > 0} \|u^{(k)}(t)\|, \quad k = 1, 2, \ldots,$$

and we assume that $u^{(k)}(t) \in L_\infty(H)$, i.e. that $D_k < \infty$. By (5) and (7) the standard estimate follows, that is,

$$\|e(t + \tau)\| = (1 + \tau L)\|e(t)\| + \tau \|R(t, u(t))\|, \quad t = 0, \tau, 2\tau, \ldots,$$

or, by recursion,
Notice that the initial and truncation errors may grow as $e^{tL}$.

By (8) we have $\|R(t, u(t))\| \leq Ct$, where $C$ depends only on the smoothness of the solution, and not on the Lipschitz constant $L$. However, in most problems of practical interest, $L$ is large, so even for moderately large values of $t$, the truncation error is amplified by a large factor $e^{-tL}$.

This is in particular true for stiff problems, in which case the bound (10) (and the method (4), even for very small steps satisfying $\tau L \ll 1$) is practically useless. This in fact holds for all explicit time-stepping methods.

However, we easily derive the following stability bound valid for solutions of the continuous problems.

$$
\|u(t) - w(t)\| \leq e^{-\int_0^t \beta(s)ds} \|u(0) - w(0)\|, \quad t > 0.
$$

Here, $u(t)$, $w(t)$ are solutions of (1) corresponding to different initial values, $u(0)$ and $w(0)$, respectively.

We now face the following problems:

(i) Can we find a numerical time-stepping method for which a similar stability bound is valid?

(ii) Can we derive discretization error estimates without a large (exponentially growing) stiffness factor, such as the factor in (10)?

The answer to these problems is affirmative as was pointed out in [4] and [5] because the backward or implicit Euler method

$$
v(t + \tau) + \tau F(t + \tau, v(t + \tau)) = v(t), \quad t = 0, \tau, 2\tau, \ldots, (11)$$

fulfills these conditions.

One finds now the error bound (if $e(0) = 0$)

$$
\|e(t)\| \leq \beta_0^{-1} \sup_{t > 0} \|R(t, u(t))\| \leq C\tau, \quad t > 0, (12)
$$

where $C$ depends only on $\beta_0$ and $D_2$.

This method is only first order accurate. In this paper we discuss an extension of (12) to the class of $\theta$-methods. The results found complement some of results in [2], [6] and [7]. In [9] necessary and sufficient conditions for Runge-Kutta methods to be contractive are presented, but under a suitable smoothness assumptions for the operator $F(t, u(t))$. 

4
2 STABILITY OF THE θ-METHOD

We shall consider the implicit (also called one-leg) form of the θ-method

\[ v(t + \tau) + \tau F(\bar{t}, \bar{v}(t)) = v(t), \quad \bar{t} = t, t + \tau, \ldots, \quad v(0) = u_0, \quad (13) \]

where \( \bar{t} = \theta(t + \tau) + (1 - \theta)t = t + \theta \tau \) and \( \bar{v}(t) = \theta v(t + \tau) + (1 - \theta)v(t) \), \( 0 \leq \theta \leq 1 \).

For \( \theta = 1 \) and \( \theta = 0 \) we get the Euler backward (i.e. the Rothe method (see [8]), for evolutionary partial differential equations) and Euler forward methods, respectively.

When the operator \( F(\cdot, \cdot) \) is monotone, i.e. satisfies (2), it follows that the nonlinear equation (13) has a unique solution \( v(t + \tau) \) in \( V \), if \( \theta \leq 1 \).

As is well-known the implicit form of the θ-method can be written as an Euler backward (implicit) step (\( \bar{t} \to \bar{t} = t + \theta \tau \))

\[ v(\bar{t}) + \tau \theta F(\bar{t}, v(\bar{t})) = v(t), \quad (14) \]

followed by an Euler forward (explicit) step (\( \bar{t} \to \bar{t} + \tau \))

\[ v(t + \tau) + \tau(1 - \theta)F(\bar{t}, v(\bar{t})) = v(t). \quad (15) \]

Equation (14) follows if we multiply (13) by \( \theta \) and define \( v(\cdot) \) as a linear function in each interval \([t, t + \tau]\), so that \( \bar{v}(t) = v(t) \). The equation (15) follows if we subtract (14) from (13).

In practice we perform errors, such as iteration and round-off errors, when solving (14) and also round-off errors when computing \( v(t + \tau) \) from (15). (In the parabolic evolution equation, we also get space discretization errors, when solving (14) and (15).) We shall assume that these errors are \( \tau r(t) \) and \( \tau s(t) \), respectively, where \( \|r(t)\| \leq C_1, \|s(t)\| \leq C_2, t \geq 0, \) and \( C_i, \quad i = 1, 2 \) are constants, independent of \( \tau \).

We get then the perturbed equations

\[ w(\bar{t}) + \tau \theta F(\bar{t}, w(\bar{t})) = w(t) + \tau r(t), \quad (16) \]

\[ w(t + \tau) + \tau(1 - \theta)F(\bar{t}, w(\bar{t})) = w(t) - \tau s(t), \quad (17) \]

which are the equations the computed solution \( w(t) \) actually satisfy.

Multiplying (16) by \( 1 - \theta \) and subtracting (17), multiplied by \( \theta \), we get

\[ w(\bar{t}) = \theta w(t + \tau) + (1 - \theta)w(t) + \tau \alpha(t) = w(t) + \tau \alpha(t), \quad (18) \]

where \( \alpha(t) = \theta s(t) + (1 - \theta)r(t) \).

By summation of (16) and (17), we find

\[ w(t + \tau) + \tau F(\bar{t}, w(\bar{t})) = w(t) + \tau \beta(t), \quad (19) \]

where \( \beta(t) = r(t) - s(t) \).

For the unperturbed equations we have

\[ v(\bar{t}) = \bar{v}(t), \quad (20) \]
respectively. Let the difference be $e(t) = w(t) - v(t)$. We find then from (18), (20) and (19), (21),
\begin{equation}
  e(t) = e(t) + \tau \alpha(t),
\end{equation}
respectively.

We shall assume that $\rho(t) \geq \rho_0 > 0$ in (2). Taking the inner product of (23) with $e(t)$, we find then, by (2) and (22),
\begin{equation}
  (e(t + \tau) - e(t), e(t) + \tau \alpha(t)) + \tau \rho_0 \|e(t) + \tau \alpha(t)\|^2 \leq \tau (\beta(t), e(t) + \tau \alpha(t)).
\end{equation}

By use of the arithmetic-geometric mean inequality, we find
\begin{equation}
  \tau (\beta(t), e(t)) \leq \frac{\tau}{2\rho_0} \|\beta(t)\|^2 + \frac{\tau \rho_0}{2} \|e(t)\|^2,
\end{equation}
and
\begin{equation}
  (e(t + \tau) - e(t), e(t) + \tau \alpha(t)) + \frac{\tau \rho_0}{4} \|e(t)\|^2 \leq \frac{1}{2} \tau \|e(t + \tau) - e(t)\|^2 + \frac{\tau}{2\rho_0} \|\beta(t)\|^2
\end{equation}
\begin{equation}
  + \tau^2 - \nu (1 + \rho_0 \tau^{1+\nu}) \|\alpha(t)\|^2,
\end{equation}
where $0 \leq \nu \leq 1$. The chosen value of $\nu$ will be specified later. See Remark 1 for motivation for this.

An elementary computation (see [2]) shows that
\begin{equation}
  (e(t + \tau) - e(t), e(t)) = \frac{1}{2} \left[ \|e(t + \tau)\|^2
\end{equation}
\begin{equation}
  + (2\theta - 1)\|e(t + \tau) - e(t)\|^2 - \|e(t)\|^2 \right]
\end{equation}
and
\begin{equation}
  \|e(t)\|^2 = \theta \|e(t + \tau)\|^2 + (1 - \theta)\|e(t)\|^2 - \theta(1 - \theta)\|e(t + \tau) - e(t)\|^2.
\end{equation}

Using these identities in (25), we find
\[ [1 + \frac{\rho_0}{2} \theta] \|e(t + \tau)\|^2 + [2\theta - 1 - \frac{\rho_0}{2} \theta(1 - \theta) - \tau'] \|e(t + \tau) - e(t)\|^2 \]
\leq [1 - \frac{\rho_0}{2} (1 - \theta)] \|e(t)\|^2 + \frac{\rho_0}{2} \|\beta(t)\|^2 + 2\tau^2 \|\alpha(t)\|^2, \quad (28)\]
where we have assumed that \( \tau \leq 1 \) is small enough so that \( \rho_0 \tau^{1+\nu} \leq 1 \), which is clearly no severe restriction for most problems. For the problems where this is a restriction, we have even a better convergence result (see Section 4 for the details).

We shall now choose \( \theta > \theta_0 \), where \( \theta_0 \) is the smallest number \( \geq 0 \), for which the factor of the second term of (28), \( 2\theta_0 - 1 - \frac{\rho_0}{2} \theta_0 (1 - \theta_0) - \tau' = 0 \). We find then \( \theta_0 = \frac{1}{2} + |O(\tau^\nu)|, \quad \tau \to 0 \), in particular \( \theta_0 \leq 1 \). By recursion, it now follows from (28),
\[ \|e(t)\|^2 \leq q^\frac{1}{2} \|e(0)\|^2 + \frac{\tau}{\rho_0} \sum_{j=1}^{(\frac{1}{2})-1} \left[ q^\frac{1}{2-j-1} \right] \left[ 1 + \frac{1}{2} \tau \rho_0 \theta \right]^{-1} \sup_{s>0} \gamma^2(s), \]
where
\[ \gamma^2(s) = \|\beta(s)\|^2 + 2\rho_0 \tau^1 \|\alpha(s)\|^2, \quad (29) \]
and
\[ q = \frac{1 - \frac{1}{2} \tau \theta - \theta (1 - \theta)}{1 + \frac{1}{2} \tau \rho_0 \theta}. \]
Since \( \theta > \frac{1}{2} \), we have \( q < 1 \), and we find
\[ \|e(t)\|^2 \leq q^\frac{1}{2} \|e(0)\|^2 + \rho_0^{-2} \left[ 2 + \tau \rho_0 \theta \right] \sup_{s>0} \gamma^2(s), \quad \forall t > 0. \]

Hence, the \( \theta \)-method is unconditionally stable (independent of the stiffness and of \( \tau \)), if \( \theta_0 \leq 1 \).

We collect the result found in

**Theorem 1 (Stability)** If (1) is strongly monotone, i.e., \( \rho(t) \geq \rho_0 \geq 0 \) in (2), and if \( \theta \geq \theta_0 \), where \( \theta_0 \) is the smallest number \( \geq 0 \), for which \( 2\theta_0 - 1 - \frac{\rho_0}{2} \theta_0 (1 - \theta_0) - \tau' = 0 \), \( 0 \leq \nu \leq 1 \), then
\[ \|e(t)\|^2 \leq q^\frac{1}{2} \|e(0)\|^2 + \rho_0^{-2} \left[ 2 + \tau \rho_0 \theta \right] \sup_{s>0} \gamma^2(s), \quad \forall t > 0, \]
where \( \gamma^2(s) \) satisfies (29). Here \( e(t) = w(t) - v(t) \) is the perturbation error, \( w(t) \) is the solution of the perturbed equations (16), (17), and \( v(t) \) is the solution of the unperturbed \( \theta \)-method (13).

**Corollary 1** If \( e(0) = 0 \), then
\[ \|e(t)\| \leq \rho_0^{-1} \left[ 2 + \tau \rho_0 \theta \right] \frac{1}{2} \sup_{s>0} \gamma(s), \forall t > 0. \quad (30) \]
This generalizes the stability part of (12) to the implicit class of \( \theta \)-methods.
3 TRUNCATION ERRORS

It remains to consider the truncation errors for the \( \theta \)-method. For the solution \( u(t) \) of (1) we have

\[
    u(t) = \tilde{u}(t) + \tau \alpha(t),
\]

where, by an elementary computation,

\[
    \alpha(t) = \theta \int_0^1 \int_s^t \int_0^s u_{tt}(t + \sigma \tau) d\sigma ds.
\]

Hence

\[
    \sup_{t>0} \| \alpha(t) \| = \frac{1}{2} \theta (1 - \theta) \tau D_2.
\]

Similarly,

\[
    u(t + \tau) + \tau F(\tilde{t}, u(\tilde{t})) = u(t) + \tau \beta(t),
\]

where

\[
    \beta(t) = \frac{1}{\tau} [u(t + \tau) - u(t) - \tau u_t(t)] = \tau \int_0^1 \int_s^t u_{tt}(t + \sigma \tau) d\sigma ds
\]

\[
    = \tau \int_0^1 \int_s^t u_{tt}(t + \sigma \tau) d\sigma ds + \tau \int_0^1 \int_s^t u_{tt}(t + \sigma \tau) d\sigma ds
\]

\[
    = \tau \int_0^1 \int_s^t u_{tt}(t + \sigma \tau) d\sigma ds + \tau \int_0^1 \int_s^t u_{tt}(t + \sigma \tau) d\sigma ds
\]

Hence, if \( u_t^{(3)}(t) \in L_{\infty}(H) \), i.e. \( \sup_{s>0} \| u_t^{(3)}(s) \| < \infty \), then

\[
    \sup_{t>0} \| \beta(t) \| \leq \frac{1}{24} \tau^2 D_3 + \tau \frac{1}{2} - \theta |D_2|.
\]

3.1 Arbitrary steps

Let the time-discretization error, \( E(t) = u(t) - v(t) \). By (20), (31) and (21), (33) and using the estimates in Section 2, we get by Corollary 1, for the strongly monotone case,

\[
    \| E(t) \| \leq \rho_0^{-1} [2 + \theta \tau \rho_0] \frac{1}{2} \sup_{t>0} |\gamma(t)|, \quad \theta_0 \leq \theta \leq 1,
\]

\[
    8
\]
where $\gamma^2(t)$ is given by (29). Hence, by (32) and (35),

$$
|\gamma(t)| = \tau \left( \frac{1}{2} - \theta \right) D_2 + \frac{1}{24} \tau^2 D_3 + \sqrt{\frac{\rho_0}{2}} \tau^{\frac{3}{2}} \theta (1 - \theta) D_2.
$$

(37)

**Remark 1** Now one can see from (29), (32) and (35) that due to the finer estimate in (25) using $\tau^\nu$ we can get a higher than 1 order of convergence. However, we still can not get optimal order of convergence, namely $O(\tau^2)$ without additional restrictions on variation of the steps. More details will be given in Section 3.2.

With $\theta = 1 - \frac{1}{2 + C \tau}, \geq \theta_0$ (i.e. with $\zeta$ a proper positive number), (37) implies

$$
|\gamma(t)| = |O(\tau^2)| + |O(\tau^{1+\nu})| + |O(\tau^{\frac{5}{2} - \nu})|, \quad \tau \to 0.
$$

The order of this bound is highest, namely $O(\tau^{\frac{3}{2}})$, if we choose $\nu = \frac{1}{3}$. We collect the result in

**Theorem 2 (Discretization error)** The discretization error of the $\theta$-method with $\theta = 1 - \frac{1}{2 + C \tau}, \geq \theta_0$, $\zeta > 0$, where $\theta_0$ is defined in Theorem 1, satisfies

$$
\|E(t)\| \leq \rho_0^{-1} \left[ 2 + \theta \tau \rho_0 \right]^{\frac{3}{2}} \sup_{t > 0} |\gamma(t)| = \left\{ \begin{array}{ll}
|O(\tau^{1+\nu})|, & \text{if} \ 0 \leq \nu \leq \frac{1}{3}, \ \forall t > 0, \\
|O(\tau^{\frac{3}{2} - \nu})|, & \text{if} \ \frac{1}{3} \leq \nu \leq 1, \ \forall t > 0,
\end{array} \right.
$$

for any solution $u(t)$ of a strongly monotone problem (1), for which $u^{(3)}(t) \in L_\infty(H)$. The order of this upper bound is highest, $\|E(t)\| = |O(\tau^{\frac{3}{2}})|$, if $\nu = \frac{1}{3}$.

**Remark 2** It readily follows from (34), that Theorem 2 remains valid if we replace the regularity requirement, $u^{(3)}(t) \in H$, with the weaker requirement that $u^{(3)}(t)$ is Hölder continuous with exponent $\nu$. In fact, it suffices that $u^{(3)}(t)$ is Hölder continuous in the interior of each interval $(t, t + \tau)$.

**Remark 3** Theorem 2 remains valid for any choice of steps $\tau_k$, constant or variable, for which $\tau_k \leq C \tau$, for some positive constant $C$.

In some problems we have to adjust the steps to get convergence or fast enough convergence, because some derivatives of $u(t)$ of low order can be discontinuous at certain points. For instance, it may happen that $F(\cdot, \cdot)$ in (1) is discontinuous for certain values of $t$.

In such cases we want to adjust the steps so that those values of $t$ become stepping-points. Hence the result in Theorem 2, although not of optimal order as we shall see, is of particular importance for cases where we have to change the steps in an irregular fashion.
3.2 Nearly constant steps

We shall now present an optimal order, $O(\tau^2)$, result, but valid only if the steps are essentially constant. The result presented here allows stronger variations of the integration steps than in [3].

Let us consider the equation (24) again. We define then

$$\delta(t) = \alpha(t) - \alpha(t - \tau), \quad t \geq 0,$$

where $\alpha(-\tau) \equiv 0$, and

$$\bar{e}(t) = e(t) + \tau \alpha(t - \tau), \quad t \geq 0,$$

Then (24) takes the following form

$$(\bar{e}(t + \tau) - \bar{e}(t) - \tau \delta(t), \bar{e}(t) + (1 - \theta)\tau \delta(t)) + \frac{\tau^2 \rho_0}{2} \|\bar{e}(t) + (1 - \theta)\tau \delta(t)\|^2 \leq \frac{\tau}{2\rho_0} \|\beta(t)\|^2.$$  (39)

By use of the inverse triangle inequality, $\|a + b\|^2 \geq \frac{1}{2}\|a\|^2 - \|b\|^2$ and the arithmetic-geometric mean inequalities we get the analog of the formula (25)

$$(\bar{e}(t + \tau) - \bar{e}(t), \bar{e}(t)) + \frac{\tau^2 \rho_0}{2} \|\bar{e}(t)\|^2 \leq \frac{1}{2}\|\bar{e}(t + \tau) - \bar{e}(t)\|^2 + \frac{\tau}{2\rho_0} \|\beta(t)\|^2 + \tau \left[ \frac{\rho_0(1-\theta)^2 \tau^2}{2} + (1 - \theta)\tau + \frac{1}{\rho_0} \right] \|\delta(t)\|^2, \quad (40)$$

Denote by $\frac{1}{2}C = \frac{\rho_0(1-\theta)^2 \tau^2}{2} + (1 - \theta)\tau + \frac{1}{\rho_0}$. Using the identities (26) and (27) in (40), we find

$$\left[ 1 + \frac{\tau^2 \rho_0}{4} \right] \|\bar{e}(t + \tau)\|^2 + \left[ 2\theta - 1 - \frac{\tau^2 \rho_0}{4}(1 - \theta) - \tau \right] \|\bar{e}(t + \tau) - \bar{e}(t)\|^2 \leq \left[ 1 - \frac{\tau^2 \rho_0}{4}(1 - \theta) \right] \|\bar{e}(t)\|^2 + \frac{\tau}{\rho_0} \|\beta(t)\|^2 + C\tau \|\delta(t)\|^2, \quad (41)$$

We shall now choose $\theta \geq \theta_0$ as above, where $\theta_0$ is the smallest number $\geq 0$, for which the factor of the second term of (41), $2\theta_0 - 1 - \frac{\tau^2 \rho_0}{4}(1 - \theta_0) - \tau = 0$. We find then $\theta_0 = \frac{1}{2} + [O(\tau)], \quad \tau \to 0$. By recursion, it now follows from (41),

$$\|\bar{e}(t)\|^2 \leq q^\frac{1}{2} \|\bar{e}(0)\|^2 + \tau \sum_{j=1}^{(\frac{1}{2})-1} \left[ q^{\frac{1}{2} - j - 1} \right] \left[ 1 + \frac{1}{4} \tau \rho_0 \theta \right]^{-1} \sup_{s > 0} \hat{\gamma}(s),$$

where

$$\hat{\gamma}(s) = \frac{1}{\rho_0} \|\beta(s)\|^2 + C\|\delta(s)\|^2, \quad (42)$$

and
\[ q = \frac{1 - \frac{1}{2} \tau \rho_0 (1 - \theta)}{1 + \frac{1}{2} \tau \rho_0 \theta}. \]

Since \( \theta > \frac{1}{2} \), we have \( q < 1 \), and we find

\[ \| \hat{\epsilon}(t) \|^2 \leq q^t \| \hat{\epsilon}(0) \|^2 + \frac{4}{\rho_0} \sup_{s > 0} \gamma^2(s), \quad \forall t > 0. \]

Hence, we proved the unconditional stability of the \( \theta \)-method (independently of the stiffness and of \( \tau \)), if \( \theta_0 \leq 1 \).

We collect the result found in

**Lemma 1** If (1) is strongly monotone, i.e. \( \rho(t) \geq \rho_0 \geq 0 \) in (2), and if \( \theta \geq \theta_0 \), where \( \theta_0 \) is the smallest number \( \geq 0 \), for which \( 2\theta_0 - 1 - \frac{\tau \rho_0}{2} \theta_0 (1 - \theta_0) - \tau = 0 \), then

\[ \| \hat{\epsilon}(t) \|^2 \leq q^t \| \hat{\epsilon}(0) \|^2 + \frac{4}{\rho_0} \sup_{s > 0} \gamma^2(s), \quad \forall t > 0, \]

where \( \gamma^2(s) \) satisfies (42). Here \( \hat{\epsilon}(t) = \epsilon(t) + \tau \alpha(t - \tau) \) is the modified perturbation error, \( \epsilon(t) \) is the perturbation error defined in Theorem 1 and \( \alpha(t) \) defined from (18).

**Corollary 2** If \( \epsilon(0) = 0 \), then by (38) we have

\[ \| \hat{\epsilon}(t) \| \leq \frac{2}{\sqrt{\rho_0}} \sup_{s > 0} | \gamma(s) |, \quad \forall t > 0. \]

For the solution \( u(t) \) of (1) we have for the variable steps \( \tau_k \) and variable weights \( \theta_k, k = 0, 1, \ldots, \)

\[ \delta(t_k) = \alpha(t_k) - \alpha(t_k - \tau_{k-1}) = \theta_k \tau_k \int_0^{1} \int_0^{\theta_k} u_{tt}(t_k + \sigma \tau_k) d\sigma ds \]

\[ - \theta_{k-1} \tau_{k-1} \int_0^{\theta_{k-1}} \int_0^{\theta_{k-1}} u_{tt}(t_k - \tau_{k-1} + \sigma \tau_{k-1}) d\sigma ds, \]

where \( t_k = \sum_{j=0}^{k-1} \tau_j \). Hence, if \( u^{(n)}(t) \in L_{\infty}(H) \),

\[ \sup_{t > 0} \| \delta(t) \| = \max_{j=1,2,\ldots} \| \tau_j \theta_j (1 - \theta_j) - \tau_j \theta_j (1 - \theta_j - 1) \| D_2 \]

\[ + \tau^2 \theta_j (1 - \theta_j - 1) D_3, \]

where \( \tau = \max_{j=0,1,\ldots, \tau_j} \).

Let the modified discretization error be defined by
Using the estimates above with obvious modifications for variable steps and weights, we get by Corollary 2, for the strongly monotone case,

$$\|E(t_k)\| \leq \frac{2}{\sqrt{\rho_0}} \sup_{t>0} |\dot{\gamma}(t)|,$$

where \(\dot{\gamma}(t)\) is given by (42). Hence, by (35) and (44),

$$|\dot{\gamma}(t_k)| = \frac{1}{\sqrt{\rho_0}} \left[ k_{1/2} - \kappa_k |D_2 + \tau_k^2 D_3\right] + \sqrt{C_{\kappa}} \left[ \tau_k \kappa_k (1 - \kappa_k) - \tau_{k-1} \kappa_{k-1} (1 - \kappa_{k-1}) |D_2 + \tau_k^2 \kappa_{k-1} (1 - \kappa_{k-1}) D_3\right],$$

With \(\theta = 1 - \frac{1}{2 + \xi \mu} \geq \theta_0\), where \(0 \leq \mu \leq 1\) (i.e. with \(\xi\) a proper positive number) and \(\tau_k \kappa_k (1 - \kappa_k) - \tau_{k-1} \kappa_{k-1} (1 - \kappa_{k-1}) = O(\tau^{1+\mu})\), (46) implies

$$|\dot{\gamma}(t)| = |O(\tau^{1+\mu})| + |O(\tau^2)| + |O(\tau^{1+\mu})| + |O(\tau^2)|, \quad \tau \to 0.$$

Its order is highest, namely \(O(\tau^2)\), if we choose \(\mu = 1\). We collect the result in

**Theorem 3** The discretization error of the \(\theta\)-method with \(\kappa_k = 1 - \frac{1}{2 + \xi_k \mu} \geq \theta_0\), \(\xi_k > 0\), \(k = 1, 2, \ldots\), where \(\xi_k\) is defined in Lemma 1, satisfies

$$\|E(t_k)\| \leq \frac{2}{\sqrt{\rho_0}} \sup_{s>0} |\dot{\gamma}(s)| + \tau \sup_{s>0} |\alpha(s)| = O(\tau^{1+\mu}),$$

where \(0 \leq \mu \leq 1\), \(\tau_k \leq \tau\), if \(\tau_k \kappa_k (1 - \kappa_k) - \tau_{k-1} \kappa_{k-1} (1 - \kappa_{k-1}) = O(\tau^{1+\mu})\), \(k = 1, 2, \ldots\), for any solution \(u(t)\) of a strongly monotone problem (1), for which \(u(t) \in L^2(H)\).

Its order is highest, \(\|E(t)\| = O(\tau^2)\), if \(\mu = 1\).

**Remark 4** Note that from the Theorem 3 it follows that the higher order of the scheme we want, the less freedom in variation of the steps we have. In particular, letting \(\kappa_k = 1 - \frac{1}{2 + \xi \mu} \geq \theta_0\), we find

$$\tau_k \kappa_k (1 - \kappa_k) - \tau_{k-1} \kappa_{k-1} (1 - \kappa_{k-1}) = \tau_k \left[ \frac{1}{4 + \xi \mu} - \frac{1}{4 + \xi_k \mu} \right] = O(\tau_k^{1+\mu})$$

which implies \(\tau_k = \tau_k + O(\tau_k^{1+\mu})\). Nevertheless, we can either increase or decrease steps.

**Remark 5** In [2], it is proven an optimal order, \(O(\tau^2)\) estimate, valid for arbitrary variable steps, if in addition to the assumptions in Theorem 2, we assume that \(\nu = 1\), that \(\|\partial F/\partial u\|\) is not large and that the Gataux derivative \(\partial F/\partial u\) exists and satisfies: \(\|\partial F/\partial u_v\|\) is of the same order as \(D_3\) (i.e. not large for smooth solutions). Note that for a linear problem \(u(t) = Au(t)\) with constant operator \(A\),
we have $\frac{\partial F(-\cdot)}{\partial u} u_{tt}(t) = A^3 u(t) = u_t^{(3)}(t)$, so this holds automatically. For a more general parabolic problem, we have typically that $\sup_{t \geq t_0} \| \frac{\partial F(-\cdot)}{\partial u} u_{tt}(t) \|$ is of the order of $\sup_{t \geq t_0} \| u_t^{(3)}(t) \|$ when the solution (and its derivatives) is smooth for $t \geq t_0$, because then $u(t)$ has essentially components along the eigenfunctions corresponding to the smallest eigenvalues of the Jacobian $\frac{\partial F(-\cdot)}{\partial u}$. In the results presented in the present paper, we have not even assumed the existence of the Jacobian.

### 4 NUMERICAL EXPERIMENTS

To illustrate the results given in the Theorems 2, 3 we make some numerical tests.

**Example 1.** The first experiment deals with the problem

$$u'(t) + \lambda (u(t) - g(t)) = g'(t), \quad t > 0, \quad \lambda \gg 1,$$

considered in [7] to show that the accuracy of the approximate solution may be unrelated to the classical order of the method used. Note that $u(t) \to g(t)$ as $t \to \infty$.

The operator $F(t, u(t)) = \lambda (u(t) - g(t)) - g'(t)$ of this problem satisfies (2) with $\rho_0 = \lambda$. In our experiment we choose $\lambda = 10^9$, $g(t) = \sin(\omega t)$, $\omega = \frac{\pi}{4}$ and $u(0) = 1.0$. Since $g(0) = 0$, there will occur an initial exponential layer in the solution. We use a variable integration step of the following type

$$\tau_k = \begin{cases} \tau, & \text{if } k \text{ is odd}, \\ 0.5\tau, & \text{if } k \text{ is even}, \end{cases} \quad k = 1, 2, \ldots, \quad (48)$$

to compute the solution of the problem $u(t) = \sin(\omega t) + \exp(-At)$ and we find the error at the time $t = 0.375$ for different values of $\theta$ and present it in the Table 1.

One can see that for the values $\theta = 0.5$ and $\theta = 0.5(1 + \tau)$ we did not get convergence at all due to the unphysical oscillations coming from the damping factor $q$ (Theorem 1). For the values $\rho_0 \tau \to \infty$ we have $q \to -\frac{1}{\omega}$. For $\theta = 0.5$ this means that $q \to -1$ as $\rho_0 \tau \to \infty$ and we have almost no damping of the oscillations, as well as in the case when $\theta = 0.5(1 + \tau)$. Now, $|q|^{\frac{1}{\tau}} \to \left(\frac{1}{1 + \tau}\right)^{\frac{1}{\tau}} \to \exp(-2t)$ as $\rho_0 \tau \to \infty$ and the oscillations decay very slowly. But for the case $\theta = 0.5(1 + \tau')$, $|q|^{\frac{1}{\tau}} \to \frac{1}{\tau'} \exp(-2\tau')$ as $\rho_0 \tau \to \infty$ and the damping is much stronger even for quite big integration steps $\tau$. We also note that the convergence rate is never less than 2 when the integration step decreases by a factor 2.

**Example 2.** In this experiment we consider again the equation (47) and we choose $\lambda = 10^9$, $g(t) = \sin(\omega t)$, $\omega = \frac{\pi}{4}$ and $u(0) = 1.0$. We use a variable integration step of the form (48) to compute the solution of the problem $u(t) = \sin(\omega t)$ and we find the error at the time $t = 9$. For this time, the initial oscillations have been damped out sufficiently. The results of the experiments are given in the Table 2.

One can easily observe the order reduction for the implicit midpoint rule ($\theta = 0.5$) that corresponds to the results obtained in [7]. On the other hand, we see that for
\[ e(t + \tau) - e(t), e(t) \Delta \leq \frac{\tau \rho_0}{4} \| e(t) \|^2 + \frac{1}{2} \tau^\nu \| e(t + \tau) - e(t) \|^2 + \frac{\tau}{2 \rho_0} \| \beta(t) \|^2 + 2 \rho_0 \tau^3 \| \alpha(t) \|^2. \] (49)

Putting \( \nu = 0 \), we arrive at the following expression for \( \gamma(s) \) instead of (29)

\[ \gamma^2(s) = \| \beta(s) \|^2 + 2 \rho_0^2 \tau^2 \| \alpha(s) \|^2. \] (50)

Hence, by (50), (32) and (35),

\[ \| \gamma(t) \| = \tau | \frac{1}{2} \tau^2 D_2 + \frac{1}{24} \tau^2 D_3 + \rho_0 \tau^2 \theta(1 - \theta) D_2. \] (51)

In spite of the fact that the first term of (51) has first order, the dominating term in the error will be the third one, as long as \( \rho_0 \tau \gg 1 \). That is why we see the second order convergence for the values of \( \theta \) far away from 0.5. The first order term becomes dominating only when \( \tau \) is so small that \( \tau \rho_0 = O(1) \).
Table 2: Example 2.

<table>
<thead>
<tr>
<th>log₂ τ</th>
<th>error for θ = 0.5</th>
<th>error for θ = 0.75</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2</td>
<td>5.4999-03</td>
<td>9.1148-05</td>
</tr>
<tr>
<td>-3</td>
<td>3.1705-03</td>
<td>5.2799-05</td>
</tr>
<tr>
<td>-4</td>
<td>1.6909-03</td>
<td>1.6688-05</td>
</tr>
<tr>
<td>-5</td>
<td>8.7180-04</td>
<td>4.5891-06</td>
</tr>
<tr>
<td>-6</td>
<td>4.4224-04</td>
<td>1.1979-06</td>
</tr>
<tr>
<td>-7</td>
<td>2.2217-04</td>
<td>3.0582-07</td>
</tr>
<tr>
<td>-8</td>
<td>1.1028-04</td>
<td>7.7236-08</td>
</tr>
<tr>
<td>-9</td>
<td>5.2832-05</td>
<td>1.9407-08</td>
</tr>
<tr>
<td>-10</td>
<td>2.1806-05</td>
<td>4.8548-09</td>
</tr>
</tbody>
</table>

Example 3. For, say, parabolic problems, where we have not only large eigenvalues but also small ones, the above exceptional result for θ > 0.5 does not hold. To model this situation we consider the following problem

$$w'(t) + F(t, w(t)) = 0, \quad t > 0, \quad w(0) = w_0,$$

where $w(t) = [u(t), v(t)]^T$ and

$$F(t, w(t)) = \frac{1}{2} \begin{bmatrix} \lambda + \mu & \mu - \lambda \\ \mu - \lambda & \lambda + \mu \end{bmatrix} \begin{bmatrix} u(t) \\ v(t) \end{bmatrix} - \frac{1}{\sqrt{2}} \begin{bmatrix} \mu g(t) + g'(t) + \lambda f(t) + f'(t) \\ \mu g(t) + g'(t) - \lambda f(t) - f'(t) \end{bmatrix}.$$

One can easily check, that $F(t, w(t))$ satisfies (2) with $\rho_0 = \min \{\lambda, \mu\}$ in the Euclidean norm and this bound is exact in the sense that it can not be improved furthermore.

Let $x(t) = 10 - (10 + t) \exp(-t)$ and $y(t) = \sin(\omega t)$. For our experiments we choose $\mu = 1$, $\lambda = 10^6$ and $w_0 = 0$. Then, $u(t) = \frac{1}{\sqrt{2}}(x(t) + y(t))$ and $v(t) = \frac{1}{\sqrt{2}}(y(t) - x(t))$ are the exact solutions of the problem. We again use the variable integration steps according to (48). The Euclidean norm of the errors at the time $t = 9$ are presented in the Table 3.
Table 3: Example 3.

One can see the order reduction for the value of $\theta = 0.5$ and the first order of convergence for fixed values $\theta$ ( here $\theta = 0.75$ ) that are bigger than 0.5. The fourth column at the Table 3 supports the theoretical results of the Theorem 2, as we can really see that the order of convergence is $\tau^{\frac{1}{2}}$. From the fifth column of the Table 3 it seems to be impossible to get second order of convergence without any restrictions on the variation of the time integration steps, as we observe the order reduction for $\theta = \frac{1}{2}(1 + \tau)$ also. However, initially for the larger integration steps, the convergence is faster for this choice. This is because there is no initial layer in the solution present at this time.

Example 4. We shall now illustrate the results stated in the Theorem 3. For this goal we take the previous example and choose the following way of variation of the integration steps

$$
\tau_k = \begin{cases} 
\tau_{k-1} - \frac{\tau^2}{2^{k-1}}, & \text{if } k \text{ is even}, \\
\tau_{k-1} + \frac{\tau^2}{2^{k-1}}, & \text{if } k \text{ is odd}, 
\end{cases}
$$

and $\tau_1 = \tau$. The results are presented in Table 4. It is seen that we have close to second order convergence.

Hence, the numerical experiments show that the method proposed in the paper
<table>
<thead>
<tr>
<th>$\log_2 \tau$</th>
<th>error for $\theta = 0.5 + \tau$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2</td>
<td>9.6709-04</td>
</tr>
<tr>
<td>-3</td>
<td>3.6468-04</td>
</tr>
<tr>
<td>-4</td>
<td>9.1431-05</td>
</tr>
<tr>
<td>-5</td>
<td>2.7713-05</td>
</tr>
<tr>
<td>-6</td>
<td>7.5457-06</td>
</tr>
<tr>
<td>-7</td>
<td>2.0341-06</td>
</tr>
<tr>
<td>-8</td>
<td>5.1210-07</td>
</tr>
<tr>
<td>-9</td>
<td>1.0938-07</td>
</tr>
<tr>
<td>-10</td>
<td>1.9286-08</td>
</tr>
</tbody>
</table>

Table 4: Example 4.

has both good rate of convergence and oscillations damping properties when solving infinitely stiff evolution equations with strongly monotone operators.

5 CONCLUSIONS

In [7] was shown by considering the problem $u'(t) + \lambda (u(t) - g(t)) = g'(t)$, $t > 0$ for $\lambda$ very large, that the accuracy of the approximate solutions obtained often are unrelated to the classical order of the method used.

For the implicit midpoint method (i.e. (13) with $\theta = \frac{1}{2}$), this error order reduction is easily seen to be caused by the damping factor $q$ in Theorem 1 approaches the value $-1$. For (nearly) constant integration steplengths this causes a cancellation effect and the global error remains $O(\tau^2)$, but for $\lambda$ and/or $\tau$ variable this is not the case and the order is only $O(\tau)$ in general, which is also seen from the numerical experiments.

We have shown that by choosing $\theta = 1 - \frac{1}{2 + \xi \nu}, \quad \zeta > 0, \quad 0 < \nu < 1$, a higher order (at least $O(\tau^{\frac{3}{2}})$) can be achieved. This is due to the damping with a factor $q$, where $|q| = \frac{1 - \theta}{\tau}$ for $\lambda \rightarrow \infty$. For $\nu = 1$, we get some damping of the initial phase term, typically, as $\exp (-\zeta \tau)$ but it is slow. For $\nu < 1$ the decay is much faster. Under an additional assumption on the ratio of the integration steps we can also get an error $O(\tau^2)$. Hence the error order is never worse than that for the implicit midpoint rule.
It is anticipated that a similar modification of higher order Lobatto type implicit Runge-Kutta methods can give a less severe order reduction than if they are not modified (see [7] and [5]).

There is an alternative method used to get a higher order of accuracy and still get a rapid damping of the initial error components. It is based on the use two values of $\theta > 0.5$ and extrapolation. For an exposition on this see [10]. The problem with this approach is that we need to solve nonlinear systems of equations two times for different values of $\theta$. This may be very expensive for a big problem. Another problem is that in [10] there is no proof for the second order of convergence of the extrapolated $\theta$-methods in general case. It is stated only that we have at least first order of convergence and we can not get more than second one.

References


