Owe Axelsson

Stability and error estimates valid for infinite time, for strongly monotone and infinitely stiff evolution equations

In: Jaromír Vosmanský and Miloš Zlámal (eds.): Equadiff 6, Proceedings of the International Conference on Differential Equations and Their Applications held in Brno, Czechoslovakia, Aug. 26 - 30, 1985. J. E. Purkyně University, Department of Mathematics, Brno, 1986.
pp. [275]--284.

Persistent URL: http://dml.cz/dmlcz/700143

STABILITY AND ERROR ESTIMATES VALID FOR INFINITE TIME, FOR STRONGLY MONOTONE AND INFINITELY STIFF EVOLUTION EQUATIONS

O. AXELSSON
Department of Mathematics, University of Nijmegen
Toernooiveld, 6525 ED Nijmegen, The Netherlands

Abstract

For evolution equation with a monotone operator we derive unconditional stability and error estimates valid for all times. For the θ-method, with $\theta = 1/(2+\zeta\tau^{\nu})$, $0<\nu\leq 1, \zeta>0$), we prove an error estimate $O(\tau^{4/3})$, $\tau \to 0$, if $\nu = 1/3$, where τ is the maximal time-step for an arbitrary choice of the sequence of timesteps and with no further condition on F, and an estimate $O(\tau^2)$ under some additional conditions. 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.

1. Introduction

Consider the evolution equation

(1.1) $\dot{u} + F(t,u) = 0$, $t > 0$, $u(0) = u_0 \in V$,

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

We shall assume that F is a monotone operator, i.e.

(1.2) $(F(t,u) - F(t,v), u-v) \geq \rho(t)\, ||u-v||^2 \quad \forall\, u, v \in V$,

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

A typical example is the parabolic evolution equation

(1.3) $u_t = \underline{\nabla} \cdot (a(t,\underline{x},\underline{\nabla}u)\underline{\nabla}u) + g(t,u)$, $t,\underline{x} \in (0,\infty) \cup \Omega$, $\Omega \subset \mathbb{R}^d$,

with boundary conditions, say $u = 0$ on $\partial\Omega$. Here $V = [\mathring{H}^1(\Omega)]^d$ (a Cauchy product of the Sobolev space $\mathring{H}^1(\Omega)$), $H = L^2(\Omega)$ and, under certain conditions on a and g this is a parabolic problem, i.e. fulfills (1.2) with $\rho(t) > 0$.

Other important examples are conservative (hyperbolic) problems for which (1.2) is satisfied with $\rho(t) \geq 0$. In the present paper we restrict the analysis to the strongly monotone case, $\rho(t) \geq \rho_0 > 0$.

Classical techniques for the derivation of discretization error estimates for (1.3), uses a semidiscrete method for the discretization in space, namely the variational form

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

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

This semidiscrete method ("longitudal 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 $C_1 h^p + C_2 \tau^q$, $p,q > 0$, where τ is the time-step. Here C_1, C_2 should be independent of h,τ. Since the dimension of the ode depends on h, classical error estimates used in the 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 (1.3), and more generally (1.1), w.r.t. time. A convenient time integration method turns out to be the implicit ("one-leg") form of the θ-method with $0 \le \theta \le 1 \,/\, 2(1+\zeta\tau^{\nu})$, $\zeta > 0$ for some ν, $0 \le \nu \le 1$. Error estimates $O((\frac{1}{2}-\theta)\tau + \tau^{(3-\nu)/2})$ valid for all t can now be derived in the strongly monotone case, where $\rho(t) \ge \rho_0 > 0$. Without further assumptions the optimal order we prove is $O(\tau^{4/3})$ for $\nu = 1/3$ and θ equal to the upper bound. With some additional assumptions we prove also the optimal order, $O(\tau^2)$. (At this point we remark that there is a marked difference in behaviour of the implicit and explicit forms of the θ-method, in particular for variable step-lengths. The latter method may not even converge.)

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

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

where v is the corresponding approximation to u.

(It is only for notational simplicity that we let the time step τ be constant.)

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

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

where V_0 contains all functions in a sufficiently large tube about the solution u. In the analysis of the Euler forward method we have to assume that F 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 (1.2). From (1.1.) it follows

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

and from (1.4) and (1.5) we get

(1.7) $\quad e(t+\tau) = e(t) - \tau\{F(t,u(t)) - F(t,v(t))\} + \tau R(t,u)$,

where

$$R(t,u) \equiv \int_0^1 [F(t,u(t)) - F(t+\tau s, u(t+\tau s))]ds = \int_0^1 [\mathring{u}(t+\tau s) - \mathring{u}(t)]ds$$

is the (normalized) local truncation error.

Note that

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

where we use the notation

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

and we assume that $u_t^{(k)} \in L_\infty(H)$, i.e. that $D_k < \infty$.

By (1.5) and (1.7) it follows

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

or, by recursion,

$$||e(t)|| \leq (1+\tau L)^{t/\tau}||e(0)|| + \tau \sum_{j=1}^{t/\tau} (1+\tau L)^{j-1}||R(t-j\tau,u)||$$

or

$$(1.10) \quad ||e(t)|| \leq e^{tL}||e(0)|| + \frac{1}{L}(e^{tL}-1) \max_{t\geq 0} ||R(t,u)||, \quad t = \tau, 2\tau, \ldots$$

Notice that the initial and truncation errors may grow as exp(tL).

By (1.8) we have $||R(t,u)|| \leq C\tau$, 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 $\sim L^{-1}\exp(tL)$.
This is in particular true for stiff problems, in which case the bound (1.10) (and the method (1.4), even for very small time-steps satisfying $\tau L \ll 1$) is practically useless. This is in fact true for all explicit time-stepping methods.

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

$$||u(t)-w(t)|| \leq \exp(\int_0^t -\rho(s)ds)\ ||u(0)-w(0)|| \leq ||u(0)-w(0)||, \quad t > 0.$$

Here, u, w are solutions of (1.1) 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 "nasty" large (exponentially growing) stiffness factor, such as the factor in (1.10)?

The answer to these problems is affirmative as was pointed out in [3] and [4] because the "backward" or implicit Euler method

$$(1.11) \quad v(t+\tau) + \tau F(t+\tau)) = v(t), \quad t = 0, \tau, 2\tau, \ldots$$

fulfills these conditions.

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

$$(1.12) \quad ||e(t)|| \leq \rho_0^{-1} \sup_{t>0} ||R(t,u)|| \leq C\tau, \quad t > 0,$$

where C depends only on ρ_0 and D_2.

This method is only first order accurate.
In this report we discuss an extension of (1.12) to the class of θ-methods. The results found complement some of the results in [2].

2. Stability of the θ-method

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

$$(2.1) \quad v(t+\tau) + \tau F(\bar{t},\bar{v}(t)) = v(t), \quad t = 0, \tau, 2\tau, \ldots,$$

$v(0) = u_0$, where $\bar{t} = \theta t + (1-\theta)(t+\tau) = t + (1-\theta)\tau$ and $\bar{v}(t) = \theta v(t) + (1-\theta)v(t+\tau)$, $0 \leq \theta \leq 1$. For $\theta = 0$ and $\theta = 1$ we get the Euler backward (i.e. the Rothe method (see [7]), for evolutionary partial differential equations and Euler forward methods, respectively.

When F is monotone, i.e. satisfies (1.2), it will follow that the nonlinear equation (2.1) has a unique solution $v(t+\tau)$ in V, if $\theta \leq 1$.

As is wellknown the implicit form of the θ-method can be written as an Euler backward (implicit) step $(t \to \bar{t} = t + (1-\theta)\tau)$.

(2.2) $v(\bar{t}) + \tau(1-\theta)F(\bar{t},v(\bar{t})) = v(t)$,

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

(2.3) $v(t+\tau) + \tau\ \theta\ k(\bar{t}) = v(\bar{t})$,

where $k(\bar{t}) = F(\bar{t},v(\bar{t}))$.

(2.2) follows if we multiply (2.1) by $(1-\theta)$ and define $v(\ )$ as a linear function in each interval $[t,t+\tau]$. Then $v(\bar{t}) = \bar{v}(t)$. (2.3) follows if we subtract (2.2) from (2.1).

In practice we perform errors, such as iteration and round-off errors when solving (2.2) and also round-off errors when computing $v(t+\tau)$ from (2.3). (In the parabolic evolution equation, we also get space discretization errors, when solving (2.2).) We shall assume that these errors are $\tau r_\theta(t)$ and $\tau s_\theta(t)$, respectively, where $||r_\theta(t)|| \leq c_1, ||s_\theta(t)|| \leq c_2$, $t \geq 0$, and c_i, $i = 1,2$ are constants, independent of τ. We get then the perturbed equations

(2.4.1) $\tilde{v}(t) + \tau(1-\theta)F(\bar{t},\tilde{v}(\bar{t})) = \tilde{v}(t) + \tau r_\theta(t)$.

(2.4.b) $\tilde{v}(t+\tau) + \tau\ \theta\ \tilde{k}(\bar{t}) = \tilde{v}(\bar{t}) - \tau s_\theta(t)$,

$\tilde{k}(\bar{t}) = F(\bar{t},\tilde{v}(\bar{t}))$,

which are the equations the computed approximations $\tilde{v}$ actually satisfy.

Multiplying (2.4.a) by θ and subtracting (2.4.b), multiplied by $(1-\theta)$, we get

(2.5) $\tilde{v}(\bar{t}) = (1-\theta)\tilde{v}(t+\tau) + \theta\tilde{v}(t) + \tau\alpha(t) = \bar{\tilde{v}}(t) + \tau\alpha(t)$,

where $\alpha(t) = \theta r_\theta(t) + (1-\theta)s_\theta(t)$.

By summation of (2.4.a) and (2.4.b), we find

(2.6) $\tilde{v}(t+\tau) + \tau F(\bar{t},\tilde{v}(\bar{t})) = \tilde{v}(t) + \tau\beta(t)$,

where $\beta(t) = r_\theta(t) - s_\theta(t)$.

For the unperturbed equations we have

(2.5') $v(\bar{t}) = \bar{v}(t)$

and

(2.6') $v(t+\tau) + \tau F(\bar{t},v(\bar{t})) = v(t)$,

respectively.

Let the difference be $e(t) = \tilde{v}(t) - v(t)$.

We find then from (2.5), (2.5') and (2.6), (2.6'),

(2.7) $e(\bar{t}) = \bar{e}(t) + \tau\alpha(t)$,

(2.8) $e(t+\tau) - e(t) + \tau[F(\bar{t},\tilde{v}(\bar{t})) - F(\bar{t},v(\bar{t}))] = \tau\beta(t)$,

respectively.

We shall assume that $\rho(t) \geq \rho(t) \geq \rho_0 > 0$ in (1.2).

Taking the inner product by (2.8) with $e(\bar{t})$, we find then, by (1.2) and (2.7),

$(e(t+\tau) - e(t), \bar{e}(t) + \tau\alpha(t)) + \tau\rho_0||\bar{e}(t) + \tau\alpha(t)||^2 \leq \tau(\beta, \bar{e}(t) + \tau\alpha(t))$.

By use of the arithmetic-geometric mean inequality, we find

$\tau(\beta, e(\bar{t})) \le \frac{1}{2}\rho_0^{-1}\tau||\beta||^2 + \frac{1}{2}\tau\rho_0||e(\bar{t})||^2$, and

$(e(t+\tau) - e(t), \bar{e}(t)) + \frac{1}{2}\tau\rho_0||\bar{e}(t) + \tau\alpha(t)||^2 \le \frac{1}{2}\rho_0^{-1}\tau||\beta||^2 - (e(t+\tau) - e(t), \tau\alpha)$.

By use of the inequality, $||a+b||^2 \ge \frac{1}{2}||a||^2 - ||b||^2$ and the arithmetic-geometric inequality once more we get

$$(2.9)\quad (e(t+\tau) - e(t), \bar{e}(t)) + \tfrac{1}{2}\tau\rho_0||\bar{e}(t)||^2 \le \tfrac{1}{2}\rho_0^{-1}\tau||\beta||^2 + \tfrac{1}{2}\tau^{\nu}||e(t+\tau)-e(t)||^2 + \tfrac{1}{2}\tau^{2-\nu}(1+\rho_0\tau^{1+\nu})||\alpha||^2,$$

where $0 \le \nu \le 1$. The chosen value of ν will be specified later.

An elementary computation (see [1] shows that

$(e(t+\tau) - e(t), \bar{e}(t)) = \frac{1}{2}[||e(t+\tau)||^2 + (1-2\theta)||e(t+\tau) - e(t)||^2 - ||e(t)||^2]$

and

$||\bar{e}(t)||^2 = (1-\theta)||e(t+\tau)||^2 + \theta||e(t)||^2 - (1-\theta)\theta||e(t+\tau) - e(t)||^2$.

Using these identities in (2.9), we find

$$(2.10)\quad [1 + \tfrac{1}{2}\tau\rho_\theta(1-\theta)]||e(t+\tau)||^2 + [1-2\theta - \tfrac{1}{2}\tau\rho_0(1-\theta)\theta - \tau^{\nu}]||e(t+\tau) - e(t)||^2 \le [1- \tfrac{1}{2}\tau\rho_0\theta]||e(t)||^2 + \rho_0^{-1}\tau||\beta||^2 + 2\tau^{2-\nu}||\alpha||^2,$$

where we have assumed that $\tau \le 1$ is small enough so that $\rho_0\tau^{1+\nu} \le 1$.

We shall now choose $\theta \le \theta_0$, where θ_0 is the largest number ≤ 1, for which the factor of the second term of (2.10), $1 - 2\theta - \frac{1}{2}\tau\rho_0(1-\theta)\theta - \tau^{\nu} \ge 0$.

We find then $\theta_0 = \frac{1}{2} - |O(\tau^{\nu})|$, $\tau \to 0$.

By recursion, it now follows from (2.10),

$$||e(t)||^2 \le q^{t/\tau}||e(0)||^2 + \tau\rho_0^{-1}\sum_{j=1}^{(t/\tau)-1} q^{(t/\tau)-j-1}[1+\tfrac{1}{2}(1-\theta)\tau\rho_0]^{-1}\sup_{s>0}\gamma^2(s)$$

where

$$(2.11)\quad \gamma^2(s) = ||\beta(s)||^2 + 2\rho_0\tau^{1-\nu}||\alpha(s)||^2,$$

and

$$q = (1 - \tfrac{1}{2}\theta\tau\rho_0) / [1 + \tfrac{1}{2}(1-\theta)\tau\rho_0]$$

Since $\theta < \frac{1}{2}$, we have $q < 1$, and we find

$||e(t)||^2 \le q^{t/\tau}||e(0)||^2 + \rho_0^{-2}[2 + (1-\theta)\tau\rho_0] \sup_{s>0}\gamma^2(s)$, $\forall t > 0$.

Hence, the θ-method is <u>unconditionally stable</u> (independent of the stiffness and of τ), if $0 \le \theta_0$.

We collect the result found in

<u>Theorem 2.1.</u> (Stability.) If (1.1) is strongly monotone, i.e. $\rho(t) \ge \rho_0 > 0$ in (1.2), and if $\theta \le \theta_0$, where θ_0 is the largest number ≤ 1, for which $1-2\theta-\frac{1}{2}\tau\rho_0(1-\theta)\theta-\tau^{\nu} \ge 0$, $0 \le \nu \le 1$, then

$||e(t)||^2 \le q^{t/\tau}||e(0)||^2 + \rho_0^{-2}[2 + (1-\theta)\tau\rho_0] \sup_{s>0}\gamma^2(s)$, $\forall t > 0$

where $\gamma(s)$ satisfies (2.11).

Here $e(t) = \tilde{v}(t) - v(t)$ is the perturbation error, $\tilde{v}(t)$ is the solution of the perturbed equations (2.4.a,b), and $v(t)$ is the solution of the unperturbed θ-method (2.1).

<u>Corollary 2.1.</u> If $e(0) = 0$, then

$$(2.12)\quad ||e(t)|| \le \rho_0^{-1}[2+(1-\theta)\tau\rho_0]^{\frac{1}{2}} \sup_{s>0}|\gamma(s)|, \quad \forall t > 0.$$

This generalizes the stability part of (1.12) to the implicit class of θ-methods.

3. Truncation errors

It remains to consider the truncation errors for the θ-method. For the solution u of (1.1) we have

$$(3.1)\quad u(\bar{t}) = \bar{u}(t) + \tau\,\alpha_\theta(t),$$

where, by an elementary computation,

$$\alpha_\theta(t) = -(1-\theta)\tau \int_0^1 ds \int_{-\theta s}^{(1-\theta)s} \ddot{u}(t-\sigma\tau)\,d\sigma$$

Hence

$$(3.2)\quad \sup_{t>0} ||\alpha_\theta(t)|| = \tfrac{1}{2}\theta(1-\theta)\tau\, D_2.$$

Similarly,

$$(3.3)\quad u(t+\tau) + \tau F(\bar{t}, u(\bar{t})) = u(t) + \tau\beta_\theta(t),$$

where,

$$(3.4)\quad \beta_\theta(t) = \tau^{-1}(u(t+\tau) - u(t) - \tau\dot{u}(\bar{t}))$$

$$= \tau \int_0^1 ds \int_{1-\theta}^{s} \ddot{u}(t+\sigma\tau)\,d\sigma$$

$$= \tau \int_0^1 ds \int_{\frac{1}{2}}^{s} \ddot{u}(t+\sigma\tau)\,d\sigma + \tau \int_0^1 ds \int_{1-\theta}^{\frac{1}{2}} \ddot{u}(t+\sigma\tau)\,d\sigma$$

$$= \tau \int_0^{\frac{1}{2}} ds \left[\int_0^s \ddot{u}(t+(\tfrac{1}{2}+\sigma)\tau)\,d\sigma - \ddot{u}(t+(\tfrac{1}{2}-\sigma))\tau\right]d\sigma + \tau \int_0^1 ds \int_{1-\theta}^{\frac{1}{2}} \ddot{u}(t+\sigma\tau)\,d\sigma.$$

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

$$(3.5)\quad \sup_{t>0} ||\beta_\theta(t)|| \le \frac{1}{24}\tau^2 D_3 + \tau|\tfrac{1}{2}-\theta|D_2.$$

Let the time-discretization error, $E(t) = u(t) - v(t)$. By (2.5'), (3.1) and (2.6'), (3.3) and using the estimates in section 2, we get by Corollary (2.1), for the strongly monotone case,

$$(3.6)\quad ||E(t)|| \le \rho_0^{-1}[2 + (1-\theta)\tau\rho_0]^{\frac{1}{2}} \sup_{t>0} |\gamma_\theta(t)|,\ 0 \le \theta \le \theta_0,$$

where $\gamma_\theta^2(t) = ||\beta_\theta(t)||^2 + 2\rho_0\tau^{1-\nu}||\sigma_\theta(t)||^2$.

Hence, by (3.2) and (3.5),

$$(3.7)\quad |\gamma_\theta(t)| = \frac{1}{24}\tau^2 D_3 + \tau|\tfrac{1}{2}-\theta|D_2 + \sqrt{\rho_0/2}\,\tau^{(3-\nu)/2}\theta(1-\theta)D_2.$$

With $\theta = 1/(2+\zeta\tau^\nu) \le \theta_0$ (i.e. with ζ a large enough positive number), (3.7) implies

$$|\gamma_\theta(t)| = |O(\tau^2)| + |O(\tau^{1+\nu})| + |O(\tau^{(3-\nu)/2})|,\ \tau \to 0.$$

Its order is highest, namely $O(\tau^{4/3})$, if we choose $\nu = 1/3$.

We collect these results in

Theorem 3.1. (Discretization error) The discretization error of the θ-method with $\theta = 1/(2+\zeta\tau^\nu) \le \theta_0$, $\zeta > 0$, where θ_0 is defined in Theorem 2.1, satisfies

$$||E(t)|| \le \rho_0^{-1}[2 + (1-\theta)\tau\rho_0]^{\frac{1}{2}} \sup_{t>0} |\gamma_\theta(t)| = \begin{cases} |O(\tau^{1+\nu})|, & \text{if } 0 \le \nu \le \frac{1}{3} \\ |O(\tau^{(3-\nu)/2})|, & \text{if } \frac{1}{3} \le \nu \le 1 \end{cases} \quad \forall\, t > 0,$$

for any solution u of a strongly monotone problem (1.1), for which $u_t^{(3)} \in L_\infty(H)$. Its order is highest, $||E(t)|| = |O(\tau^{4/3})|$, if $\nu = 1/3$.

Remark 3.1. It follows readily from (3.4), that Theorem 3.1 remains valid if we replace the regularity requirement, $u_t^{(3)} \in H$, with the weaker requirement that $u_t^{(2)}$ is Hölder-continuous with exponent ν. In fact it suffices that $u_t^{(2)}$ is Hölder-continuous in the interior of each interval $(t,t+\tau)$.

Remark 3.2. Theorem 3.1 remains valid for any choice of timesteps τ_k, constant or variable, for which $\tau_k \leq C\tau$, for some positive constant C.

In some problems we have to adjust the timesteps to get convergence or fast enough convergence, because some derivative of u of low order can be discontinuous at certain points. It may for instance happen that F in (1.1) is discontinuous for certain values of t.

In such cases we want to adjust the timesteps so that those values of t become stepping-points. Hence the result in Theorem 3.1, although not of optimal order as we shall see, is of particular importance for cases where we have to change the time-steps in an irregular fashion.

We shall now present an optimal order, $O(\tau^2)$, result, but valid only if the timesteps are essentially constant.

Consider first the equations (2.4.a,b) for variable parameters $\theta = \theta_k$ and $\tau = \tau_k$, $k = 0,1,\ldots$. For the solution of (1.1) we get then truncation errors $R_k = R(t_k,\tau_k,\theta_k)$ and $S_k = S(t_k,\tau_k,\theta_k)$, defined by

(3.8.a) $u(\bar{t}) + \tau_k(1-\theta_k)F(\bar{t},u(\bar{t})) = u(t_k) - \tau_k R_k$,

(3.8.b) $u(t_{k+1}) + \tau_k\theta_k F(\bar{t},u(\bar{t})) = u(\bar{t}) + \tau_k S_k$,

where $\bar{t} = t_k + (1-\theta_k)\tau_k$ and t_k is the k'th stepping point.

Guided by a trick in [5] for the implicit midpoint method (i.e. (2.2), (2.3) with $\theta = \frac{1}{2}$) we define

(3.9) $\hat{u}(t_k) = u(t_k) - \tau_k R_k$.

Then (3.8.a,b) takes the form

(3.10.a) $u(\bar{t}) + \tau_k(1-\theta_k)F(\bar{t},u(\bar{t})) = \hat{u}(t_k)$,

(3.10.b) $\hat{u}(t_{k+1}) + \tau_k\theta_k F(\bar{t},u(\bar{t}) = u(\bar{t}) + \tau_k\hat{\beta}_\theta(t_k)$,

where

(3.11) $\hat{\beta}_\theta(t_k) = S_k - \tau_{k+1}R_{k+1}/\tau_k$

Multiplying (3.10.a) by θ_k and subtracting (3.10.b), multiplied by $(1-\theta_k)$, yields

(3.12.a) $u(\bar{t}) = \theta_k\hat{u}(t_k) + (1-\theta_k) + (1-\theta_k)\hat{u}(t_{k+1}) - (1-\theta_k)\hat{\beta}_\theta(t_k)$

Summation of (3.10.a) and (3.10.b) yields

(3.12.b) $\hat{u}(t_{k+1}) + \tau_k F(\bar{t},u(\bar{t})) = \hat{u}(t_k) + \tau_k\hat{\beta}_\theta t_k)$.

We define $\hat{u}(\bar{t}) = u(\bar{t})$ (we let $0 < \theta_k$, so $t_k < \bar{t} < t_{k+1}$), so by (3.12.a),

(3.13) $\hat{u}(t) = \bar{\hat{u}}(t_k) + \tau_k\hat{\alpha}_\theta(t_k)$, where $\hat{\alpha}_\theta(t_k) = -(1-\theta_k)\hat{\beta}_\theta(t_k)$.

Note that (3.13) and (3.12.b) have the same form as (3.1) and (3.3), respectively.

To estimate $\hat{\alpha}_\theta$ and $\hat{\beta}_\theta$, we need to estimate R_k and S_k. By (3.8.a) and (3.8.b) we find

$$(3.14)\quad R_k = (1-\theta_k)\,[\mathring{u}(\bar{t}) - \int_0^1 \mathring{u}(t_k + (1-\theta_k)\tau_k s)ds]$$
$$= (1-\theta_k)^2\tau_k \int_0^1 ds \int_s^1 \ddot{u}(t_k + (1-\theta_k)\tau_k\sigma(d\sigma)$$

and

$$(3.15)\quad S_k = \theta_k[\int_0^1 \mathring{u}(t_k+\tau_k-\theta_k\tau_k s)ds - \mathring{u}(\bar{t})]$$
$$= \theta_k^2\tau_k \int_0^1 ds \int_s^1 \ddot{u}(t_k+\tau_k-\theta_k\tau_k\sigma)d\sigma.$$

By (3.11), (3.14) and (3.15) we find

$$||\hat{\beta}_\theta(t_k)|| = \tau_k^{-1}|(\theta_k\tau_k)^2 - (1-\theta_{k+1})^2\tau_{k+1}^2|\,\tfrac{1}{2}D_2 + O(\tau_k^2)D_3$$

if $c\tau \le \tau_k \le C\tau$ for some positive constants c,C.

Hence $||\hat{\beta}_\theta(t_k)|| = O(\tau^2)$, $\tau \to 0$, $k = 0,1,\ldots$ if $(\theta_k\tau_k)^2 - (1-\theta_{k+1})^2\tau_{k+1}^2 = O(\tau^3)$ or

$$\tau_{k+1} = \frac{\theta_k}{1-\theta_{k+1}}\tau_k + O(\tau^2).$$

Remark 3.3. Since for stability reason, $\theta_k < \frac{1}{2}$, we see that if

$$(3.16)\quad \tau_{k+1} = \frac{\theta_k}{1-\theta_{k+1}}\tau_k,$$

then $\{\tau_k\}$ is a decreasing sequence. Frequently, in practice we want to choose small steps in the initial (transient period) and then larger steps, i.e. contrary to (3.16).

Consider now for simplicity the case $\tau_k = \tau$, $\theta_k = \theta = 1/(2+\zeta\tau)$, $\zeta > 0$, $k = 0,1,\ldots$. Let $\hat{E}(t) = \hat{u}(t) - v(t)$. It follows as before (see 3.6) and Theorem 2.1) that

$$||\hat{E}(t)||^2 \le q^{t/\tau}||\hat{E}(0)||^2 + \rho_0^{-2}\,[2 + (1-\theta)\tau\rho_0]\,O(\tau^2)D_3$$

We have $\hat{E}(0) = \hat{u}(0) - v(0) = \hat{u}(0) - u(0) = \tau_0 R_0$, i.e. by (3.9) and (3.14), $||\hat{E}(0)|| = O(\tau^2)D_2$.

Similarly, by (3.1),

$$||u(t) - v(t)|| \le ||u(t) - \hat{u}(t)|| + ||\hat{u}(t) - v(t)|| = O(\tau^2)D_2 + ||\hat{E}(t)||.$$

We collect these results in

Theorem 3.2. If $\tau_{k+1} = \frac{\theta_k}{1-\theta_{k+1}}\tau_k + O(\tau^2)$,

$\theta_k = 1/(2+\zeta_k\tau_k)$, $\zeta > \zeta_k > 0$, $k = 0,1,\ldots$, then the θ-method (2.2), (2.3) has a discretization error $O(\tau^2)$, valid at all stepping points t_k, if $u_t^{(3)} \in L_\infty(H)$ and if (1.1) is strongly monotone.

Remark 3.4. In [2], it is proven an optimal order, $O(\tau^2)$ estimate, valid for arbitrary variable time-steps, if in addition to the assumptions in Theorem 3.1, we assume that $\nu = 1$, that $||\partial F/\partial t||$ is not large and that the Gataux derivative $\partial F/\partial u$ exists and satisfies: $||\partial F/\partial u\, u_t^{(2)}||$ is of the same order as D_3 (i.e. not large for smooth solutions). Note that for a linear problem $u_t = Au$ with constant operator A, we have $\partial F/\partial u\, u_t^{(2)} = A^3 u = u_t^{(3)}$. For a more general parabolic problem, we have typically that $\sup_{t\ge t_0} ||\partial F/\partial u\, u_t^{(2)}||$ is of the order of $\sup_{t\ge t_0} ||u_t^{(3)}||$ when the solution (and its

derivatives) is smooth for $t \geq t_0$, because then u has essentially components along the eigenfunctions corresponding to the smallest eigenvalues of the Jacobian $\partial F/\partial u$. In the results presented in the present paper, we have however not even assumed the existence of the Jacobian.

4. Conclusions

In [6] was shown by considering the problem $\mathring{u}(t) + \lambda(u-g(t)) = \mathring{g}_{(t)}$, $t > 0$ for λ 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. (2.1) with $\theta = \frac{1}{2}$), this error order reduction is easily seen to be caused by that the damping factor q in Theorem 2.1 approaches the value -1. For (almost) constant steplength this causes a cancellation effect and the global error remains $O(\tau^2)$, but for λ and/or τ variable this is not the case and the order is only $O(\tau)$ in general.

We have shown that by choosing $\theta = 1/(2+\zeta\tau^{\nu})$, $\zeta > 0$, $0 < \nu < 1$, a higher order (at least $O(\tau^{4/3})$) can be achieved. This is due to the damping with a factor q, where $|q| \simeq \theta/(1-\theta)$ for λ large.
Under additional assumptions and with $\nu = 1$ we can also get an error $O(\tau^2)$. Hence the error order is never worse 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 (cf [6] and [4]).

References

1. O. Axelsson, Error estimates for Galerkin methods for quasilinear parabolic and elliptic differential equations in divergence form, Numer. Math. 28, 1-14(1977).

2. O. Axelsson, Error estimates over infinite intervals of some discretizations of evolution equations, BIT 24(1984), 413-424.

3. G. Dahlquist, Error analysis for a class of methods for stiff nonlinear initial value problems, Numerical Analysis (G.A. Watson, ed.), Dundee 1975, Springer-Verlag, LNM 506, 1976.

4. R. Frank, J. Schneid and C.W. Ueberhuber, The concept of B-convergence, SIAM J. Numer. Anal. 18(1981), 753-780.

5. J. Kraaijevanger, B-convergence of the implicit midpoint rule and the trapezoidal rule, Report no. 01-1985, Institute of Applied Mathematics and Computer Science, University of Leiden, The Netherlands.

6. A. Prothero and A. Robinson, The stability and accuracy of one-step methods, Math. Comp. 28(1974), 145-162.

7. K. Rektorys, The Method of Discretization in Time and Partial Differential Equations. D. Reidel Publ. Co., Dordrecht-Holland, Boston-U.S.A., 1982.