Extrapolated $\theta$-methods for nonlinear reaction-diffusion problems. *

O.AXELSSON † S.V.GOLOLOBOV † YU.M.LAEVSKY §

Abstract

The present paper discusses the solution of semilinear parabolic problems of the form: $u_t(t,\mathbf{x}) = L(t, \mathbf{x})$, $t > 0$, $\mathbf{x} \in \Omega$, with standard boundary conditions and initial condition $u(0) = u_0$. Here $L(t, \mathbf{x}) = -\nabla \cdot (k(t, \mathbf{x}) \nabla u(t, \mathbf{x})) + g(t, u(t, \mathbf{x}))$, where $f(t, u(t, \mathbf{x}))$ corresponds to a nonlinear reaction term. It is assumed that the solution has a transient initial phase, where it has big gradients. Because of the unboundedness of the operator $L(e, \cdot)$ we use an implicit numerical integration method. As an efficient such solution method an implicit form of the extrapolated $\theta$-method is used combined with simple mesh size and time step controls. For proper values of $\theta$, this method has good stability properties, gives high order accurate approximations and requires solution only of relatively simple linear or nonlinear systems. Each arising nonlinear system, which is nonsymmetric in general, is solved by a modified Newton method. Some numerical tests are presented that illustrate the results obtained.

1 INTRODUCTION

Consider the parabolic problem

$$u_t(t, \mathbf{x}) + L(t, u(t, \mathbf{x})) = 0, \quad t > 0, \quad \mathbf{x} \in \Omega,$$

with standard boundary conditions and initial condition $u(0, \mathbf{x}) = u_0(\mathbf{x})$. Here $L(t, u(t, \mathbf{x})) = -\nabla \cdot (k(t, \mathbf{x}) \nabla u(t, \mathbf{x})) + g(t, u(t, \mathbf{x}))$, where $f(t, u(t, \mathbf{x}))$ corresponds to a nonlinear reaction term. It is assumed that the solution has a transient initial phase, where big gradients of the solution occur. Such a behavior is typical for this equation because the initial solution does not satisfy the differential equation $\lim_{t \to 0^+} u(t, \mathbf{x}) + L(t, u(t, \mathbf{x})) = 0$ and the problem is stiff, i.e. the operator $L(e, \cdot)$ is unbounded.

---

*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.

§Institute of Computational Mathematics and Mathematical Geophysics, Siberian Branch of Russian Academy of Sciences, Lavrentiev av. 6, 630090 Novosibirsk, Russia. E-mail: laev@labchem.sscc.ru.
The purpose of the paper is to describe efficient solution methods which give good resolution of the solution of the PDE.

As the solution has an initial transient phase, the efficiency of the method depends strongly on a good mesh size control. One of the first who have started with the adaptive strategies for parabolic equations was T. Dupont [5]. In his approach only the points are moved in the mesh, but its number is not changed. Much of investigations based on the discontinuous Galerkin approximation in time (backward Euler approximation corresponds to the discontinuous Galerkin method of order 0, where the integral in the right-hand side is approximated by some low order quadrature) was made by K. Eriksson and C. Johnson et al. [6-8, 11]. Some refinement strategies for the linear parabolic problems were proposed by R.E. Ewing, R.D. Lazarov and A.T. Vassilev in [9-10].

Because of the unboundedness of the operator $L(\cdot, \cdot)$ an implicit time-discretization method will be used. An efficient class of such methods are the $\theta$-methods. However, the order of accuracy is limited by order one, or order two for the Crank-Nicolson version of the method. Extrapolation combining the solution for different values of $\theta$ may possibly be used to improve the accuracy. In Section 2 will be shown that there does not exist an extrapolated $\theta$-method of order higher than two neither for the "static" nor for the "dynamic" extrapolation strategy. The terms "static" and "dynamic" extrapolation strategy mean that we don't use the extrapolated value of the function in further calculations for the former and do this for the latter. The alternative of getting of high order of accuracy using Richardson extrapolation methods ($\tau, \frac{\tau}{2}, \ldots$) is less efficient when solving problems with transient behavior and therefore, are not further considered in the paper.

During the numerical solution each arising nonlinear system, which is nonsymmetric in general, will be solved by a modified Newton method. This will be considered in Section 4. Some tests for a population ecology model will be described, for which we obtain good practical results in combination with an adaptive mesh refinement method. At the end some conclusions are drawn.

2 TIME-DISCRETIZATION METHOD

2.1 The $\theta$-method

The $\theta$-method is a method which generalizes the Euler forward and Euler backward methods as well as the trapezoidal method. When solving a differential equation

$$u_t + L(t,u) = 0, \quad t > 0, \quad u(0) = u_0.$$  \hfill (1)

it takes the following form: let $v(t)$ be a piecewise linear function, i.e. linear on each interval $(t, t + \tau)$, which satisfies

$$v(t + \tau) + \tau L(t, \tilde{v}(t)) = v(t), \quad t = \tau, 2\tau, \ldots,$$  \hfill (2)
where $\bar{t} = t + \theta \tau$, $\bar{v}(t) = \theta v(t + \tau) + (1 - \theta) v(t)$. For proper values of the method parameter the method has good stability properties. For instance, as shown in [3], if $L(\cdot, \cdot)$ is monotone

$$(L(t, u(t)) - L(t, v(t)), u(t) - v(t)) \geq 0, \quad \forall t \in \mathbb{R}^+, \quad \forall u, v,$$

then the method satisfies the following stability estimate:

$$||v_1(t) - v_2(t)|| \leq ||v_1(0) - v_2(0)||,$$

The discretization error estimate of the method is readily derived and takes the form

$$||e(t)|| = O(\tau^2) + |1 - 2\theta| O(\tau),$$

where $e(t) = u(t) - v(t)$. This shows that it is only of first order of approximation unless $\theta = \frac{1}{\zeta}(1 + \zeta \tau)$ for some constant $\zeta$ in which case it is of second order. For $\zeta = 0$ we get the trapezoidal (also called the implicit form of the Crank-Nicolson) method.

However, the damping of the initial transient phase is insufficient in this method and can cause oscillations leading to unphysical triggerings of the nonlinear reaction term. This occurs because for $\lambda \gg 1$ the coefficient for the eigenvector corresponding to the eigenvalue $\lambda$ is multiplied at each step by a factor $\frac{1 - \theta \lambda}{1 + \theta \tau \lambda} < 0$. It is somewhat better for values $\theta = \frac{1}{\zeta}(1 + \zeta \tau)$, $\zeta > 0$, for which the second order of convergence still holds, but as the asymptotic decay factor in the latter case is

$$\left( \frac{1 - (1 - \theta) \tau \lambda}{1 + \theta \tau \lambda} \right)^\frac{\tau}{\theta} \to \left( \frac{1 - \theta \lambda}{\theta} \right)^{\frac{\tau}{\theta}} \sim \left( \frac{1 - \zeta \tau}{1 + \zeta \tau} \right)^{\frac{\tau}{\theta}} \sim e^{-2\zeta \tau},$$

it is still too slow.

Therefore it is of interest to consider values of $\theta$ which give much better damping. To still retain a higher order of accuracy, one can then combine the approximations for two or more values of $\theta$ and extrapolate the results to find a method of higher order of approximation.

Using $m$ values of $\theta$ within each time-step, it may seem plausible that one can derive a method of $m$-th order of accuracy in the time-step parameter. This will be next.

### 2.2 Extrapolated $\theta$-methods for a model problem

We consider first the following model problem:

$$u_t + \lambda u(t) = 0, \quad t > 0, \quad u(0) = u_0 \in \mathbb{R},$$

where $u_t = \frac{du}{dt}$, $\text{Re}(\lambda) > 0$. Then we get a dissipative equation. In [3] was shown the stability property for dissipative systems in particular of the $\theta$-method on implicit form for $\text{Re}(\lambda) \geq 0$, for $t \geq 0$. Let $m$ be a positive integer. Choose $\theta_i$ such that $\frac{1}{\tau} \leq \theta_i \leq 1$, $i = 1, \ldots, m$. First, we consider the extrapolated $\theta$-method on implicit
form with "static" extrapolation (step (3) is an implicit Euler, while step (4) is a forward Euler explicit step) and it can be readily seen that method (2) can be rewritten on this form:

\[
v_i(\tilde{t}_i) + \tau \theta_i \lambda v_i(\tilde{t}_i) = v_i(t), \quad (3)
\]

\[
v_i(t + \tau) + \tau(1 - \theta_i) \lambda v_i(\tilde{t}_i) = v_i(\tilde{t}_i), \quad (4)
\]

\[
v(t + \tau) = \sum_{i=1}^{m} \phi_i v_i(t + \tau), \quad (5)
\]

\[
v_i(0) = v_i, \quad (6)
\]

where \( \tilde{t}_i = t + \theta_i \tau, \quad v_i(t) = \theta_i v_i(t + \tau) + (1 - \theta_i)v_i(t), \quad \phi_i \in \mathbb{R}, \quad i = 1, \ldots, m \) and where the extrapolation parameters \( \phi_i, \quad i = 1, \ldots, m \) are to be determined. As was shown in [3] each of the processes of computations of \( v_i(t) \) is asymptotically stable. The proof can be found in Subsection 2.3 for the general case. Let us consider now the extrapolated \( \theta \)-method on implicit form with "dynamic" extrapolation:

\[
v_i(\tilde{t}_i) + \tau \theta_i \lambda v_i(\tilde{t}_i) = v(t), \quad (7)
\]

\[
v_i(t + \tau) + \tau(1 - \theta_i) \lambda v_i(\tilde{t}_i) = v_i(\tilde{t}_i), \quad (8)
\]

\[
v(t + \tau) = \sum_{i=1}^{m} \phi_i v_i(t + \tau), \quad (9)
\]

\[
v_i(0) = v_i, \quad (10)
\]

One can see that in both extrapolated strategies each \( v_i(t), \quad i = 1, \ldots, m \), can be computed in parallel with perfect load balance. Using the fact that \( u_i(\tilde{t}_i) = -\lambda u(\tilde{t}_i) \) we get for the local truncation errors the following expression (the details for the general case will be given in the Subsection 2.3):

\[
R_{\theta}(t, u) = \alpha_{\theta}(t) - \beta_{\theta}(t), \quad (11)
\]

where

\[
\alpha_{\theta}(t) = \sum_{i=1}^{m} \phi_i \left[ \frac{u(t + \tau) - u(t)}{\tau} - u_i'(\tilde{t}_i) \right],
\]

and

\[
\beta_{\theta}(t) = \sum_{i=1}^{m} \phi_i \lambda (u_i(t) - u(\tilde{t}_i)).
\]

**Lemma 1** Assume that for all \( t > 0, \quad u(t) \in C^{m+1}(0, \infty) \). Then

\[
\alpha_{\theta}(t) = \sum_{k=0}^{m-1} \frac{u^{(k+1)}(t)}{k!} \left[ \frac{1}{k+1} - \sum_{i=1}^{m} \phi_i \theta_i^k \right] + O(\tau^m).
\]
Proof. The statement of the Lemma follows using elementary Taylor series expansion.

Similarly, we get

**Lemma 2** Assume that for all \( t > 0, \ u(t) \in C^m(0, \infty). \) Then

\[
\beta_i(t) = \sum_{i=1}^{m} \phi_i \lambda \sum_{l=2}^{m-1} u^{(l)}(t) \frac{t^l}{l!} (\theta_l - \theta_i^l) + O(t^m).
\]

**Theorem 1** Independently of the choice of parameters \( \phi_i, \ i = 1, \ldots, m \) and integer \( m > 0 \) neither a "static" nor a "dynamic" extrapolated \( \theta \)-method of order three exists.

Proof. From the Lemmas 1, 2 it follows that we must satisfy four equations concurrently to get an extrapolated \( \theta \)-method of the third order uniformly on \( \lambda. \) From Lemma 1 we obtain the following three equations at \( k = 0, 1, 2: \)

\[
\sum_{i=1}^{m} \phi_i = 1, \tag{12}
\]

\[
\sum_{i=1}^{m} \phi_i \theta_i = \frac{1}{2}, \tag{13}
\]

\[
\sum_{i=1}^{m} \phi_i \theta_i^2 = \frac{1}{3}, \tag{14}
\]

On the other hand, from the Lemma 2 we get the equation for \( l = 2: \)

\[
\sum_{i=1}^{m} \phi_i (\theta_i - \theta_i^2) = 0.
\]

This equation contradicts the equations (13), (14) above. As the local truncation error satisfies the equation (11) the Theorem is proved.

2.3 The "statically" extrapolated \( \theta \)-method for the reaction-diffusion problems

Consider now the following problem:

\[ u_t + L(t, u) = 0, \quad t > 0, \quad u(0) = u_0 \in V, \]

where \( V \) is a reflexive Banach space, where \( u_t = \frac{\partial u}{\partial t} \) and \( L(t, \cdot) : V \to V'. \) Here \( V' \) is the space which is dual with respect to the inner product \((\cdot, \cdot)\) in a Hilbert space \( H, \) with norm \( (v, v) = ||v||^2. \) We shall assume that operator \( L(\cdot, \cdot) \) satisfies the condition

\[
(L(t, u) - L(t, v), u - v) \geq \rho(t)||u - v||^2, \quad \forall t \in \mathbb{R}^+, \ \forall u, v \in V, \tag{15}
\]
where $\rho(t) : (0, \infty) \to \mathbb{R}$. In [2] was shown in particular the stability property of the $\theta$-method on implicit form for $\rho(t) \geq \rho_0 > 0$, for $t \geq 0$. For this case ( diffusion type problems ) we get the best stability property. For it (see [2]) we get a stable $\theta$-method for $\frac{1}{2} \leq \theta \leq 1$.

Consider the extrapolated $\theta$-method on implicit form with "static" extrapolation:

\begin{align}
   v_i(\tilde{t}_i) + \tau \theta_i L(\tilde{t}_i, v_i(\tilde{t}_i)) &= v_i(t), \quad (16) \\
   v_i(t + \tau) + \tau (1 - \theta_i) L(\tilde{t}_i, v_i(\tilde{t}_i)) &= v_i(\tilde{t}_i), \quad (17) \\
   v(t + \tau) &= \sum_{i=1}^{m} \phi_i v_i(t + \tau), \quad (18) \\
   v_i(0) &= v_0, \quad (19)
\end{align}

where $\tilde{t}_i = t + \theta_i \tau$, $v_i(\tilde{t}_i) = \tilde{v}_i(t) = \theta_i v_i(t + \tau) + (1 - \theta_i) v_i(t)$, $\phi_i \in \mathbb{R}$, $i = 1, 2$.

Repeating the proofs in [2] we get the stability estimates like in [2]. Let us consider the perturbed equations

\begin{align}
   \tilde{\tilde{v}}_i(\tilde{t}_i) + \tau \theta_i L(\tilde{t}_i, \tilde{\tilde{v}}_i(\tilde{t}_i)) &= \tilde{\tilde{v}}_i(t) + \tau \tau_i(t), \\
   \tilde{\tilde{v}}_i(t + \tau) + \tau (1 - \theta_i) L(\tilde{t}_i, \tilde{\tilde{v}}_i(\tilde{t}_i)) &= \tilde{\tilde{v}}_i(\tilde{t}_i) - \tau s_i(t),
\end{align}

which are the equations the computed approximation $\tilde{v}_i(t)$ actually satisfy. Here $r_i(t)$, $s_i(t)$ can include round-off errors and errors due to not solving the, in general, nonlinear equation exactly and they are assumed to satisfy $\|r_i(t)\|, \|s_i(t)\| \leq const$. From them we get

\begin{align}
   \tilde{\tilde{v}}_i(\tilde{t}_i) &= \tilde{\tilde{v}}_i(t) + \tau \alpha_i(t), \quad (20) \\
   \tilde{\tilde{v}}_i(t + \tau) + \tau L(\tilde{t}_i, \tilde{\tilde{v}}_i(\tilde{t}_i)) &= \tilde{\tilde{v}}_i(\tilde{t}_i) + \tau \beta_i(t), \quad (21)
\end{align}

where

\begin{align}
   \alpha_i(t) &= (1 - \theta_i) r_i(t) + \theta_i s_i(t), \\
   \beta_i(t) &= r_i(t) - s_i(t).
\end{align}

Let the difference be $e_i(t) = \tilde{v}_i(t) - v_i(t)$. we shall now analyze the stability of the $\theta$-method with respect to these perturbation errors. We find then from (16), (17), (20), (21)

\begin{align}
   e_i(\tilde{t}_i) &= e_i(t) + \tau \alpha_i(t), \quad (22) \\
   e_i(t + \tau) - e_i(t) + \tau [L(\tilde{t}_i, \tilde{v}_i(\tilde{t}_i)) - L(\tilde{t}_i, v_i(\tilde{t}_i))] &= \tau \beta_i(t). \quad (23)
\end{align}

Taking the inner product of (23) with $e_i(\tilde{t}_i)$, we find then, by (15) and (22)
\[
(e_i(t + \tau) - e_i(t), e_i(t) + \tau \alpha_i(t)) + \tau \rho(t) \| e_i(t) + \tau \alpha_i(t) \|^2 \\
\leq \tau (\beta_i(t), e_i(t) + \tau \alpha_i(t)).
\]

By using of the arithmetic-geometric mean inequalities with the weights \(e, \frac{1}{\tau}\) and \(\tau^\nu\), where \(0 \leq \nu \leq 1\), respectively we get

\[
(e_i(t + \tau) - e_i(t), e_i(t)) + \frac{\tau}{\nu} (\rho(t) - \frac{\tau}{\nu}) \| e_i(t) \|^2 \leq \frac{1}{\nu} \tau^\nu \| e_i(t + \tau) - e_i(t) \|^2 \\
+ \frac{\tau}{\nu} \| \beta_i(t) \|^2 + \tau^{2-\nu} \left[ \frac{1}{2} + (\rho(t) - \frac{\tau}{\nu}) \tau^{1+\nu} \right] \| \alpha_i(t) \|^2,
\]

where \(\epsilon > 0\) is some constant, say \(\epsilon = |\rho_0|\). An elementary computations (see [2]) shows that

\[
\left[ 1 + \tau (\rho(t) - \frac{\tau}{\nu}) \theta_i \right] \| e_i(t + \tau) \|^2 \\
+ \left[ 2 \theta_i - 1 - \tau (\rho(t) - \frac{\tau}{\nu}) \theta_i (1 - \theta_i) - \tau^\nu \right] \| e_i(t + \tau) - e_i(t) \|^2 \\
\leq \left[ 1 - \tau (\rho(t) - \frac{\tau}{\nu}) (1 - \theta_i) \right] \| e_i(t) \|^2 \\
+ \frac{\tau}{\nu} \| \beta_i(t) \|^2 + 2\tau^{2-\nu} \| \alpha_i(t) \|^2,
\]

(24)

where we have assumed that \(\tau \leq 1\) is small enough so that \((\rho(t) - \frac{\tau}{\nu}) \tau^{1+\nu} \leq \frac{1}{2}\) for all times \(t\). We shall now choose \(\theta_i \geq \theta_0\), where \(\theta_0\) is the largest number \(\leq 1\), for which the factor of the second term of (24) \(2 \theta_0 - 1 - \tau (\rho(t) - \frac{\tau}{\nu}) \theta_0 (1 - \theta_0) - \tau^\nu = 0\). We find then \(\theta_0 = \frac{1}{2} + |O(\tau^\nu)|, \tau \to 0\). By recursion, it now follows from (24)

\[
\| e_i(t) \|^2 \leq \prod_{k=1}^{\frac{t}{\tau}} q_k(k\tau) \| e_i(0) \|^2 \\
+ \frac{\tau}{\nu} \sum_{j=1}^{\frac{t}{\tau} - 1} \left[ \prod_{k=1}^{\frac{t}{\tau} - j - 1} q_k(k\tau) \right] \left[ 1 + \tau (\rho(j\tau) - \frac{\tau}{\nu}) \theta_i \right]^{-1} \sup_{p > 0} \gamma_i^2(p),
\]

(25)

where

\[
\gamma_i^2(p) = \| \beta_i(p) \|^2 + 2\tau^{1-\nu} \| \alpha_i(p) \|^2
\]

and

\[
q_i(t) = \frac{1 - \tau (\rho(t) - \frac{\tau}{\nu}) (1 - \theta_i)}{1 + \tau (\rho(t) - \frac{\tau}{\nu}) \theta_i}.
\]

Note, that if \(\rho(t) \geq \rho_0 > 0\), \(\forall t > 0\) and as \(\theta_i > \frac{1}{2}\) then we may take \(\epsilon = \rho_0\) and get

\[
q_i(t) \leq q_{i,0} = \frac{1 - \frac{\tau}{\nu} \rho_0 (1 - \theta_i)}{1 + \frac{\tau}{\nu} \rho_0 \theta_i} < 1
\]

and find

\[
\| e_i(t) \|^2 \leq q_{i,0}^2 \| e(0) \|^2 + \rho_0^{-2} [2 + \tau \rho_0 \theta_i] \sup_{p > 0} \gamma(p), \forall t > 0.
\]

7
Moreover, if \( \forall t \geq t_0 \), where \( t_0 \) is some fixed time, we have \( \rho(t) \geq \rho_0 > 0 \) then we again get the 0 stability of the process of the computations of \( v_i(t) \). Note, that this estimate (25) means that \( \rho(t) \) may be even \( \leq 0 \) for a finite period of time then we still have the asymptotic stability of the method.

For the local truncation error we get an expression like (11):

\[ R_\theta(t, u) = \alpha_\theta(t) - \beta_\theta(t), \tag{26} \]

where

\[ \alpha_\theta(t) = \sum_{i=1}^{m} \phi_i \left[ \frac{u(t + \tau) - u(t)}{\tau} - u_i^{(1)}(\tilde{t}_i) \right] \]

and

\[ \beta_\theta(t) = \sum_{i=1}^{m} \phi_i \left( L(\tilde{t}_i, \bar{u}_i(t)) - L(\tilde{t}_i, u(\tilde{t}_i)) \right). \]

For \( \alpha_\theta(t) \) we have already an estimate. For \( \beta_\theta(t) \) the next Lemma can be used.

**Lemma 3** Assume that for all \( t > 0 \), \( u(t) \in C^2(V) \) and the inequality

\[ \|L(t, u) - L(t, v)\| \leq K\|u - v\|, \quad \forall u, v \in V, \tag{27} \]

holds. Then

\[ \|\beta_\theta(t)\| \leq \frac{\tau^2}{2} K\|u^{(2)}(t)\| \sum_{i=1}^{m} |\phi_i| (\theta_i - \theta_i^2). \]

**Proof.** Using the inequality (27) and Lemma 2 we get the statement of the Lemma.

We can now use Lemmas 1, 3 to estimate also the discretization error from the truncation errors.

**Theorem 2** For the diffusion type problems with the operator \( L(t, u) \) satisfied the condition (15) with \( \rho(t) \geq \rho_0 > 0 \) there exists a stable "statically" extrapolated \( \theta \)-method in implicit form of order not higher than two for all times \( T \).

**Proof.** From the Lemma 1 it follows that we must satisfy two equations concurrently to get the extrapolated \( \theta \)-method of the second order:

\[ \sum_{i=1}^{m} \phi_i = 1, \quad \tag{28} \]

\[ \sum_{i=1}^{m} \phi_i \theta_i = \frac{1}{2}. \quad \tag{29} \]
Let $m = 2$. In our case $\theta_1, \theta_2$ assumed to be greater than $\theta_0 > \frac{1}{2}$. Thus, if $\theta_1 \neq \theta_2$, their unique solution is

$$\phi_1 = \frac{\theta_2 - \frac{1}{2}}{\theta_2 - \theta_1} \quad \text{and} \quad \phi_2 = \frac{\frac{1}{2} - \theta_1}{\theta_2 - \theta_1}.$$  

Hence, if we use the ”static” extrapolation method with $\frac{1}{2} < \theta_0 \leq \theta_1, \theta_2 \leq 1$ and $\theta_1 \neq \theta_2$ then, according to the [2], we get a stable extrapolated $\theta$-method. As the local truncation error satisfies the equation (26) the Theorem is proved.

We have used a uniform mesh in time. But all results hold for an arbitrary non-uniform mesh. Note, that for $\frac{1}{2} < \theta_1, \theta_2 \leq 1$ the coefficient $\phi_2$ becomes negative. But this does not destroy the stability of the method as we don’t use the extrapolated value in the further computations. In the initial transient phase it is recommended to take values of $\theta_i, i = 1, 2$, close to 1 to damp out the components in the error corresponding to the largest eigenvalues. For further comments on this see e.g. [4]. However, one shall not chose the values of $\theta_1, \theta_2$ very close to each other as this causes the growth of the absolute values of the weight coefficients $\phi_1, \phi_2$ and, hence, the growth of the round-off error. In order to avoid this situation it is recommended to take the coefficients $\theta_1, \theta_2$ in such a way that the weights $\phi_1, \phi_2$ by absolute value will not be higher than the computed solution. If even for this choice of the parameters $\theta_1, \theta_2$ the damping of the high frequency components of the solution is still 0, it seems good to use the ”dynamic” extrapolation strategy with special choice of the parameters, where both $\theta_1, \theta_2$ may be greater than 1.

### 2.4 The ”dynamically” extrapolated $\theta$-method for monotone operators

Consider now the ”dynamic” approach. For this approach we have the following extrapolated $\theta$-method:

$$v_i(t_i) + \tau \theta_i L(\tilde{t}_i, v_i(t_i)) = v(t),$$
$$v_i(t + \tau) + \tau(1 - \theta_i) L(\tilde{t}_i, v_i(t_i)) = v_i(t_i),$$
$$v(t + \tau) = \sum_{i=1}^{m} \phi_i v_i(t + \tau),$$
$$v_i(0) = v_0,$$

where $\tilde{t}_i = t + \theta_i \tau, \; v_i(t_i) = \tilde{v}_i(t) = \theta_i v_i(t + \tau) + (1 - \theta_i) v_i(t), \; \phi_i \in \mathbb{R}, \; i = 1, \ldots, m$. The difference with the ”static” approach is only in the first equation, where we take the value of the extrapolated function $v(t)$ for all $m$ equations instead of different values $v_i(t), \; i = 1, \ldots, m$. We assume now that the operator $L(\cdot, \cdot)$ satisfies the condition (15) with $\rho(t) > 0$. For the local truncation error we get an expression like (26). Hence, we can conclude from the Theorem 1 that there does not exist an extrapolated $\theta$-method of order higher than two.
Now we shall investigate the stability properties of this method for the model problem only. Then one can see that the equations (7)-(9) can be written in the following form

\[ v(t + \tau) = P(\tau \lambda) v(t), \]  

(34)

where \( P(\tau \lambda) = \phi_1 \frac{1-(1-\theta_1)\tau\lambda}{1+(\theta_1+\theta_2)\tau\lambda} + \phi_2 \frac{1-(1-\theta_2)\tau\lambda}{1+(\theta_1+\theta_2)\tau\lambda}, \) or in the equivalent form using the unique solution of (28), (29) for \( m = 2 \)

\[ P(\tau \lambda) = \frac{1 + \left(\theta_1 + \theta_2 - 1\right)\tau\lambda + \left(\frac{1}{\tau} - \theta_1 - \theta_2 + \theta_1 \theta_2\right)\tau^2 \lambda^2}{1 + \left(\theta_1 + \theta_2\right)\tau\lambda + \theta_1 \theta_2 \tau^2 \lambda^2} \]  

(35)

**Lemma 4** a). If \( \frac{1}{\tau} \leq \theta_1 + \theta_2 \leq \frac{1}{\tau} + 2\theta_1 \theta_2, \ \theta_1 \geq 0, \ \theta_2 \geq 0, \) then the method (7)-(10) is \( A \)-stable.

b). Moreover, if in addition \( \theta_1 + \theta_2 \geq 1 \) and \( \theta_1 + \theta_2 \leq \frac{1}{\tau} + \theta_1 \theta_2 \) then the operator \( P(\tau \lambda) > 0 \) for all \( \lambda \geq 0. \)

**Proof.** a). From the conditions of the Lemma 4 it immediately follows that \( |P(\tau \lambda)| \leq 1 \) for any \( \lambda \in \mathbb{C} \) such that \( \text{Re}(\lambda) \geq 0. \)

b). Under the additional conditions one can easily check that \( P(\tau \lambda) > 0 \) for any \( \lambda \in \mathbb{R} \) such that \( \lambda \geq 0. \)

**Lemma 5** a). If \( \theta_2 = \frac{\theta_1 - \theta_1}{2}, \ \theta_1 > 0, \ \theta_2 > 0, \) it holds for the transition operator of the method (7)-(10) \( P(\tau \lambda) \to 0, \) when \( \lambda \to +\infty, \) that is, the method is \( L \)-stable.

b). Moreover, if in addition \( \theta_1 + \theta_2 \geq 1 \) then the operator \( P(\tau \lambda) > 0 \) for all \( \lambda \geq 0. \)

**Proof.** The proof is similar to that of Lemma 4.

It is clear, that the "dynamic" extrapolation can also be considered on an arbitrary non-uniform mesh with the same results.

Hence, we conclude:

- for both the "dynamic" and the "static" extrapolation strategies there does not exist an extrapolated \( \theta \)-method of order higher than two

- the "static" extrapolation with \( m = 2 \) is asymptotically stable for any time \( T \) for diffusion type problems

- the "dynamic" extrapolation with \( m = 2 \) is \( A \)-stable ( \( L \)-stable ) under some restrictions on the parameters \( \theta_1 \) and \( \theta_2. \) Moreover, the transition operator can be made positive for all eigenvalues. This property is very important in order to avoid unphysical triggerings in the computed solution, see [4].

### 3 Refinement Criteria

Now, we consider the criteria for control of the space mesh and the time step.
3.1 Space mesh refinement criterion

Let us consider the original equation

\[ u_t + L(t, u) = 0, \quad t > 0, \text{ in a domain } \Omega, \]

with standard boundary conditions and initial condition \( u(0) = u_0 \). Apply to it the method of lines, that is, we first discretize this equation with respect to space variables. Now we look for the solution \( \tilde{u}(t) \) of the following system of ordinary differential equations:

\[ \tilde{u}_t(t) + \tilde{L}(t, \tilde{u}(t)) = 0, \quad t > 0, \text{ in a domain } \Omega_h, \]

with discretized boundary conditions and initial condition \( \tilde{u}(0) = \tilde{u}_0 \). Here \( \tilde{L}(\cdot, \cdot) \) is a discrete operator corresponding to the operator \( L(\cdot, \cdot) \). For the sake of simplicity we assume that we approximate the boundary conditions with the required accuracy. Then the local truncation error satisfies the following equation:

\[ u_t(t) + \tilde{L}(t, u(t)) = R(t), \quad t > 0. \]

Let \( e(t) = u(t) - \tilde{u}(t), \) then

\[ e_t(t) + \tilde{L}(t, u(t)) - \tilde{L}(t, \tilde{u}(t)) = R(t). \]

Multiply the last equation by \( e(t) \). Assume now that the operator \( \tilde{L}(\cdot, \cdot) \) is also strongly monotone like \( L(\cdot, \cdot) \), but with a constant \( \tilde{\rho}_0 \). It’s quite natural requirement that the discrete operator preserves the monotonicity property of the continuous one. Thus, using the strong monotonicity property of the operator \( \tilde{L}(\cdot, \cdot) \) and the arithmetic-geometric mean inequality one finds that

\[ \frac{d}{dt} ||e(t)||^2 + \tilde{\rho}_0 ||e(t)||^2 \leq \frac{1}{\tilde{\rho}_0} ||R(t)||^2. \]

Now, we multiply this inequality by \( e^{\rho_0 t} \) and after integrating the resulting equation one gets

\[ ||e(t)||^2 \leq \exp(-\tilde{\rho}_0 t)||e(0)||^2 + \frac{\exp(-\tilde{\rho}_0 t)}{\tilde{\rho}_0} \int_0^t ||R(t)||^2 \exp(\tilde{\rho}_0 s) ds. \]

If we make \( ||R(t)|| \leq TOL_x \) then

\[ ||e(t)||^2 \leq \exp(-\tilde{\rho}_0 t)||e(0)||^2 + \frac{TOL_x^2}{\tilde{\rho}_0}. \] (36)

Note that the time discrete analog of the inequality (36) one can find in [2] for \( \theta \)-method. In our further experiments we always have that \( e(0) = 0 \).
3.2 Time-step control

Let us consider again the equation

\[ \ddot{u}_i(t) + \dot{L}(t, \ddot{u}(t)) = 0, \quad t > 0, \quad \text{in a domain } \Omega_h, \]

with discretized boundary conditions and initial condition \( \ddot{u}(0) = \ddot{u}_0 \). Using the same proof as in [2] for the strongly monotone operator \( L(\cdot, \cdot) \) one can get the estimate for the \( \| \ddot{e}(t) \| \) of the same form as in [2], but with the constant \( \rho_0 \) replaced by \( \tilde{\rho}_0 \), where \( \ddot{e} = \ddot{u} - u \), \( \ddot{u} \) is the solution for the \( \theta \)-method applied to the equation (37). Under the assumption that the initial error is zero one can prove at least for the "static" extrapolation, that

\[ \| \ddot{u}(t) - \ddot{u}_i(t) \| \leq C_t(t + h^2), \]

where \( \ddot{u}(t) = \phi_1 \ddot{u}_1(t) + \phi_2 \ddot{u}_2(t) \) and \( \ddot{u}_i(t) \) is the approximate solution for (37) with \( \theta = \theta_i, \quad i = 1, 2 \). Here and at latter occurrences \( C_j, \quad j = 1, 2, \ldots \), stands for the constant independent on \( \tau \) and \( h \). Assume that \( h \leq C_3 \tau \) and

\[ \| \ddot{e}(t) \| \leq C_4 \tau^2 + C_5 h^2 \leq C_6 \tau^2. \]

Hence, for the control of the error we can use the following inequality

\[ \| \ddot{e}(t) \| \leq C_7 \cdot TOL_t, \]

where \( TOL_t \geq \tau \| \ddot{u}(t) - \ddot{u}_i(t) \| \). In our experiments we choose \( i = 1 \).

4 NUMERICAL EXPERIMENTS

Consider now a practical problem – the population ecology model. The next system of nonlinear elliptic equations was proposed to model certain planktonic predator-prey situations in which crowding is a factor of importance. For further details, see [12].

Let \( v(x, t) \) denote the number of predators, i.e. zooplankton, and \( u(x, t) \) the essentially static number of prey, i.e. phytoplankton. Then the system reads in the domain \( \Omega = (0, 2.5) \) for \( t > 0 \):

\[
\begin{align*}
  u_t - 0.0125 \ u_{xx} &= \left( \frac{35 + 16u - u^2}{9} - v \right) u \\
  v_t - v_{xx} &= \left( u - \frac{5 + 2v}{5} \right) v
\end{align*}
\]

with the boundary conditions

\[ u_x = v_x = 0 \text{ for } t > 0, x = 0 \text{ and } x = 2.5 \]

12
and initial conditions

\[
u_0 = \begin{cases} 
5 & \text{for } 0 \leq x < 1.0, \\
4x + 1 & \text{for } 1.0 \leq x < 1.25, \\
-4x + 11 & \text{for } 1.25 \leq x < 1.5, \\
5 & \text{for } 1.5 \leq x < 2.5. 
\end{cases}
\]

\[
v_0 = \begin{cases} 
10 & \text{for } 0 \leq x < 1.0, \\
4x + 6 & \text{for } 1.0 \leq x < 1.25, \\
-4x + 16 & \text{for } 1.25 \leq x < 1.5, \\
10 & \text{for } 1.5 \leq x < 2.5. 
\end{cases}
\]

Let us check now the condition (15) for the nonlinear operator of this problem in the \(L_2\)-inner product, that is consider \((\mathcal{L}(\overline{w}) - \mathcal{L}((\overline{\overline{w}}), w - \overline{w}), \mathcal{w} = (u, v)^T\).

For this purpose we shall consider first the operator

\[
E(\xi) = -\left[ \begin{array}{c} \frac{3\xi_1^2 - \xi_2^2}{9} - \xi_2 \\ \xi_1 - \frac{5\xi_2}{9} \end{array} \right] \xi_1.
\]

Direct computations show that the Frechet derivative of this nonlinear mapping is

\[
E'(\xi) = -\left[ \begin{array}{cc} \frac{3\xi_2^2 - 3\xi_2}{9} - \xi_2 & -\xi_1 \\ \frac{\xi_2}{9} & \left( \xi_1 - \frac{5\xi_2}{9} \right) \end{array} \right].
\]

For the eigenvalues of the derivative we get the following quadratic equation

\[
\lambda^2 - Tr(-E'(\xi))\lambda + det(-E'(\xi)) = 0.
\]

Now, we observe that the vector \(\xi = [5 - 5 \cos \frac{6\pi}{2.5} x, 10 - \cos \frac{6\pi}{2.5} x]^T\) is quite close to the computed steady state solution and, hence, we may expect that it is close to the exact solution as well. Let us denote \(y = \cos \frac{6\pi}{2.5} x\), thus, \(y \in [-1, 1]\). As a result we get that

\[
-\mathcal{E}'(\xi) = \left[ \begin{array}{cc} \frac{30y - 75y^2}{10 - y} & 5y - \frac{5}{9} \\ 30y - 21y & 5y - \frac{5}{9} \end{array} \right]
\]

And one can easily check that

\[
-45 \det(-\mathcal{E}'(\xi)) \equiv P(y) = 1575y^3 + 1746y^2 - 3105y + 1650.
\]

Now, let \(y \in [0, 1]\) and \(Q(y) = 1746y^2 - 3105y + 1650\), then \(P(y) \geq Q(y) \geq \min Q(y) > 50\). For the case when \(y \in [-1, 0]\) we take \(Q(y) = 1575y^3 + 1650\) and again \(P(y) \geq Q(y) \geq \min Q(y) > 75\). Hence,

\[
\det(-\mathcal{E}'(\xi)) > 1 > 0.
\]

13
For the case, when $\lambda_1, \lambda_2 \in \mathbb{R}$ this inequality means in particular, that $\text{sign}(\lambda_1) = \text{sign}(\lambda_2)$.

Further, we find that

$$45 \, \text{Tr}(-E'(\xi)) \equiv P(y) = -375y^2 - 194y - 30.$$  

It is easy to see, that $P(y) < \max P(y) < -3$, $\forall y \in [-1, 1]$, that is,

$$\text{Tr}(-E'(\xi)) < \frac{-1}{15} < 0.$$  

As

$$\det(-E'(\xi)) = \lambda_1 \lambda_2 \quad \text{and} \quad \text{Tr}(-E'(\xi)) = \lambda_1 + \lambda_2,$$

we get that $Re(\lambda_1), Re(\lambda_2) < 0$, which implies that the operator $E'(\xi)$ is positive semi-definite. This can be rewritten as follows, using continuous dependence of the eigenvalues of the matrix from the entries of this matrix,

$$(F(w)) - F(\hat{w}), w - \hat{w}) \geq 0, \quad \forall w, \hat{w} \in B_R(\xi),$$

where $B_R(\xi)$ is some ball around the vector $\xi$. Assuming that $||\xi - \hat{u}|| < R$, where $\hat{u}$ is the exact steady state solution of the ecology problem, we may find a small ball $B_r(\hat{u})$, where the operator $E'(\cdot)$ is nonnegative. As a result we get that the operator of the problem

$$L(\hat{u}) = \left[ \begin{array}{c} -0.0125 (u_1)_{xx} \\ -(u_2)_{xx} \end{array} \right] + F(\hat{u})$$

satisfies the inequality (15) with $\rho(t) \geq 0.0125(\frac{\pi}{2\pi})^2$ in the ball $B_r(\hat{u})$, where $\hat{u}$ is the exact steady state solution of the problem. Here we have used the well-known fact about the smallest eigenvalue of the 1-dimensional Laplace operator.

We also can find the numerical estimate for the function $\rho(t)$ at the time $t \approx 10$. For this purpose we use the $\theta$-method with "static" extrapolation. Let the vectors $\hat{w}, \hat{w}_r$ correspond to the solution computed by $\theta$-method with $\theta = \theta_1$ and $\theta = \theta_2$ respectively. Using the well-known result about the smallest eigenvalue of the Laplace operator and the explicit formula for the operator $L(\cdot)$ we get for $t \approx 10$

$$\rho \approx \frac{(L(\hat{w}) - L(\hat{w}_r), w - \hat{w})}{||w - \hat{w}||^2} \approx 3.7.$$  

Thus, we may expect that near the stationary solution, that is starting from the time $t \geq 10$ we have the estimate (15) with $\rho(t) \geq 3$. Thus the numerical experiment shows that our analytical estimate for the parameter $\rho(t)$ is very rough.

In all further experiments we assume that the absolute tolerance is equal to 0.01 for the modified Newton method to compute the intermediate solutions $u_i(\bar{t}_{n,i})$ and $v_i(\bar{t}_{n,i}), \quad i = 1, 2, \quad n = 1, 2, ...$. This tolerance is good enough to get practical results.
We shall assume that the "steady state" was reached after 10 time units. As was mentioned in [12], this is a reasonable choice of the "steady state" time.

According to the results of the Subsection 3.1 we add a new point in the mesh in the interval \((x_{i-1}, x_i), \ i = 2, N + 1\), at the time \(t_n\) if

\[
\frac{h^2}{12} \max \left[|u_x^{(4)}(t_n, x_{i-1})|, |u_x^{(4)}(t_n, x_i)|\right] \geq \frac{TOL_x}{\sqrt{N + 1}},
\]

and we remove the point \(x_i\) from the mesh if

\[
\frac{h^2}{12} \max \left[|u_x^{(4)}(t_n, x_{i-1})|, |u_x^{(4)}(t_n, x_i)|, |u_x^{(4)}(t_n, x_{i+1})|\right] \leq 0.25 \frac{TOL_x}{\sqrt{N + 1}}.
\]

Then the discrete \(L_2\)-norm of the space error will not be greater than \(TOL_x\). Note, that for this problem it is essential the preservation of the symmetry of the refinement process. To save computational efforts, we compute \(u_x^{(4)}(t_n, x_i)\) only in the points corresponding to the initial uniform grid with 10 points and with the accuracy \(h = 0.25\) and then interpolate these values on the intermediate mesh points. Also we keep the discrete \(L_2\)-norm of the space error between 70\% and 90\% of the initial value \(TOL_x\).

A new time step \(\tau_n\) is computed, using the following formula

\[
\tau_n = \tau_{n-1} \frac{\sqrt{\|u(t_{n-1}) - u_1(t_{n-1})\|^2 + \|v(t_{n-1}) - v_1(t_{n-1})\|^2}}{\sqrt{\|u(t_n) - u_1(t_n)\|^2 + \|v(t_n) - v_1(t_n)\|^2}},
\]

and we keep

\[
\tau_n \sqrt{\|u(t_n) - u_1(t_n)\|^2 + \|v(t_n) - v_1(t_n)\|^2} \leq TOL_t,
\]

where \(TOL_t = 0.001\). Then the discrete \(L_\infty\) norm over the time and the discrete \(L_2\)-norm over the space of the error will be less than \(C_7 \cdot TOL_t\). In addition to this control we use the restriction that

\[
\tau_n \leq 2 \cdot \tau_{n-1},
\]

to preserve the stability of the computations. As before, by increasing of the time step \(\tau_n\) twice, we don't allow \(\tau_n \{\|u(t_n) - u_1(t_n)\|^2 + \|v(t_n) - v_1(t_n)\|^2\}^{1/2}\) to be less than \(0.2 \cdot TOL_t\).

Consider also the influence of the modified Newton iterations on the convergence behavior of the solution.

It is easy to see that we have the following expression for the discretization error in the modified Newton method:

\[
\|e_0\| = \|w_i(\tilde{t}) - w_i(t)\| = O(\tau + h^2), \quad i = 1, 2.
\]

Hence, for small enough parameters \(\tau\) and \(h\) we get that the Newton method converges, because the initial approximation \(w_i(t), \ i = 1, 2\), will be in the interior of
the domain of convergence ( see [13] ). This immediately implies that after the first iteration we get the following estimate for the error \( \|e_1\| \):

\[
\|e_1\| = O(\tau^2 + h^4).
\]

That is, the error of the modified Newton iteration becomes of the order of the discretization error after the first iteration. A few additional number of modified Newton iterations make the error of it of order less than the order of the discretization error. Therefore the error arising from the modified Newton method has no influence on the order of accuracy of the extrapolated \( \theta \)-methods.

To solve the arising linear system of equations in the modified Newton method we use the standard MATLAB function, that corresponds to the Gaussian elimination method. We choose the following values of the parameters: \( \theta_1 = 0.6, \ \theta_2 = 0.9 \) for the "static" extrapolation method and special values \( \theta_1 = 2.1, \ \theta_2 = 1.5 \) ( note that they both are greater than 1 ) for the "dynamic" one. The development of \( u \)- and \( v \)-components of the solution in the former case are on the Figure 1 and for the latter one can be seen on the Figure 2.

The "static" approach shows a more smooth behavior with respect to the time-stepping procedure than the "dynamic" one. We obtain the evolution of grid points as in the Figure 3, where the grid was developed from 10 points up to 302 for both "static" and "dynamic" extrapolation strategies. Because of the rapid change of the solution in the initial transient stage more modified Newton iterations are needed there to get the norm of the residual less than 0.01. In this case we must do approximately 3-4 modified Newton iterations in comparison with 2 iterations at the end of the time interval.

The results of the experiments are presented in the Table 1. Note that time-step for the uniform mesh is even smaller than the initial time-step for both extrapolation strategies. This occurs due to the fact that the sharpest gradients of the solution are situated near the time \( t = 3.5 \).

<table>
<thead>
<tr>
<th>Experiment</th>
<th>&quot;static&quot;</th>
<th>&quot;dynamic&quot;</th>
<th>uniform</th>
</tr>
</thead>
<tbody>
<tr>
<td>initial ( N )</td>
<td>10</td>
<td>10</td>
<td>324</td>
</tr>
<tr>
<td>final ( N )</td>
<td>300</td>
<td>302</td>
<td>324</td>
</tr>
<tr>
<td>initial ( \tau )</td>
<td>0.059</td>
<td>0.059</td>
<td>0.016</td>
</tr>
<tr>
<td>final ( \tau )</td>
<td>( \approx 0.84 )</td>
<td>( \approx 1.19 )</td>
<td>0.016</td>
</tr>
</tbody>
</table>

Table 1: Results of the experiments.
Evolution of the \textit{u}-component.

Evolution of the \textit{v}-component.

Figure 1: "Static" extrapolation.
Evolution of the u-component.

Evolution of the v-component.

Figure 2: "Dynamic" extrapolation.
5 CONCLUSIONS

In this paper the possibility of getting an extrapolated $\theta$-method of higher order has been considered. It was found that neither the "static" nor the "dynamic" extrapolation strategy can not give an extrapolated $\theta$-method of order higher than two. This was proved in the Theorem 1. The proposed methods are perfectly parallelizable on two processors. Both extrapolation strategies are equally parallelizable, but the "dynamic" one requires somewhat more communications. Our experiments show good stability and approximation properties for the population ecology model with initial transient phase for both types of extrapolated $\theta$-method. Using simple time step control we can get a significant gain in computations with respect to the use of a uniform mesh. For this particular test example we don’t get a big gain from the use of the space mesh-size control. However, it can be significant for solutions which are smooth almost everywhere except in a small part of the domain, where they have big gradients such as in a boundary layer. The "dynamic" extrapolation approach shows better stability and convergence behavior than the "static" one when the parameters
${\theta}_i, \ i = 1, 2,$ are greater than 1. But the "static" approach shows more smooth behavior with respect to the timestep control. The extrapolated $\theta$-method yields less accurate solutions than implicit Runge-Kutta methods, but can be faster when one is contact with accuracies of orders magnitude frequently used in practice.

6 ACKNOWLEDGMENTS

We would like to thank Dr. M.Netycheva and M.Nikolova (both at the Catholic University of Nijmegen, The Netherlands) for their help in the development of refinement strategies.

References


