CWI

Centrum voor Wiskunde en Informatica
Centre for Mathematics and Computer Science

W.H. Hundsdorfer

Stability results for *θ*-methods applied to a class of stiff differential-algebraic equations

Department of Numerical Mathematics Report NM-R8708 April

θ-

Stability Results for θ-Methods Applied to a Class of Stiff Differential-Algebraic Equations

W.H. Hundsdorfer
Centre for Mathematics and Computer Science
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

In this paper we consider some simple numerical methods for a class of stiff differential-algebraic equations (with index 2). The methods are based on the well known θ-method for ordinary differential equations. The stability and some convergence properties of the methods are discussed.

1980 Mathematics Subject Classification: 65L05, 65L20.
Key Words & Phrases: differential-algebraic equations, stiff initial value problems.

1. Introduction

1.1. The class of differential-algebraic equations

We consider the following system of differential-algebraic equations (DAEs)

$$\dot{v}(t) = F(t,v(t)) - Aw(t), \tag{1.1a}$$

$$0 = B(v(t)+g(t)) \tag{1.1b}$$

where $v(t) \in \mathbb{R}^{m_1}$ and $w(t) \in \mathbb{R}^{m_2}$ are unknown, and $0 \leqslant t \leqslant 1$. Further $F\colon \mathbb{R}\times\mathbb{R}^{m_1} \to \mathbb{R}^{m_1}$, $g\colon \mathbb{R} \to \mathbb{R}^{m_1}$ and the linear operators $A \in L(\mathbb{R}^{m_2},\mathbb{R}^{m_1})$, $B \in L(\mathbb{R}^{m_1},\mathbb{R}^{m_2})$ are given, together with an initial value v^0 in $\mathbb{R}^{m_1}$,

$$v(0) = v^0. \tag{1.2}$$

We assume $m_1 \geqslant m_2$ and

$$BA \textit{ is regular.} \tag{1.3}$$

Systems of the type (1.1) arise for instance as semi-discrete (space discretized) versions of the Navier-Stokes equations for incompressible fluids, in which case v represents the velocity field and w the pressure field, which is fixed in some given point of the spatial domain. The boundary values are then incorporated in F and g, and F may also contain forcing terms. The DAE system (1.1) is a so called index 2 system. For nonstiff problems of this type convergence results for BDF methods were obtained by Lötstedt and Petzold [7], and Gear, Leimkuhler and Gupta [5]. Here we will consider stiff systems, but we confine ourselves to a small class of methods.

When using a standard ODE method, like BDF, for the numerical solution of (1.1), we usually compute at each time level new approximations to v and w simultaneously, see [5], [7]. An alternative, that will be considered in section 3, is to first compute a prediction to v by an ODE method, using only (1.1a) and freezing the $Aw(t)$ term, and then projecting this prediction onto the plane defined by the algebraic constraints (1.1b). This results in a scheme where the computation of v and w can be done successively, which reduces the dimension of the algebraic equations to be solved at each step. For the Navier-Stokes equations such procedures were introduced by Chorin [1] and Temam [12].

The main object of this paper is to analyze to what extent the stability properties of the original ODE method are affected by such a prediction-projection procedure. We shall restrict our attention to

Report NM-R8708
Centre for Mathematics and Computer Science
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

a simple one-step method, the θ-one leg method, as the underlying ODE method. The stability analysis of this method itself is relatively simple. Further it will be assumed that the DAE system (1.1) satisfies certain stability requirements. These requirements do allow the system to be arbitrarily stiff, no bounds are imposed on the Lipschitz constant of the function F, nor on the norms of the matrices A and B.

1.2. Stability of the differential-algebraic equations

In the following it will be assumed that F and g are continuously differentiable. Let the projection C in $\mathbb{R}^{m_1}$ be defined by

$$C = A(BA)^{-1}B. \tag{1.4}$$

By differentiation of the algebraic constraints (1.1b) it follows from (1.1a) that

$$\dot{v}(t) = (I-C)F(t,v(t)) - C\dot{g}(t), \tag{1.5a}$$

$$Aw(t) = CF(t,v(t)) + C\dot{g}(t). \tag{1.5b}$$

The system (1.5) in its turn implies that (1.1a) holds and that $(d/dt)B(v(t)+g(t)) = 0$, so that with a consistent initial value for v, $0 = B(v^0+g(0))$, we reobtain (1.1b). Thus (1.1) is equivalent with (1.5), provided that the initial value is consistent with the algebraic constraints.

It is convenient to consider $Aw(t)$ as dependent variable instead of $w(t)$. If $Aw(t)$ is known we can always compute $w(t)$ from $w(t) = [(BA)^{-1}B]Aw(t)$. The matrix $I-C$ arising in (1.5) is a projection on the plane $\{u \in \mathbb{R}^{m_1} : Bu = 0\}$. If $A^T = B$ this projection is orthogonal w.r.t. the Euclidian inner product.

The Euclidian inner product on the spaces $\mathbb{R}^m, m \geqslant 1$, will be denoted by (x,y), and $|x| = (x,x)^{1/2}$ is the corresponding norm. For any function G: $\mathbb{R}^m \rightarrow \mathbb{R}^m$ we denote by $\mu[G]$ its one-sided Lipschitz constant and by $\|G\|$ its Lipschitz constant,

$$\mu[G] = sup\{(Gx-Gy,x-y)/|x-y|^2 : x,y \in \mathbb{R}^m, x \neq y\},$$

$$\|G\| = sup\{|Gx-Gy|/|x-y| : x,y \in \mathbb{R}^m, x \neq y\}.$$

(Usually the names logarithmic norm and spectral norm are used for $\mu[G], \|G\|$, if G is linear).

It will be assumed in the rest of this paper that there are constants $\alpha,\beta,\gamma \geqslant 0$ such that for any $t \in [0,1]$

$$\|CF(t,\cdot)\| \leqslant \alpha, \ \mu[(I-C)F(t,\cdot)] \leqslant \beta, \ \|C\| \leqslant \gamma. \tag{1.6}$$

These assumptions imply that the system (1.1) is stable in the following sense. Consider beside (1.1) a perturbed version

$$\dot{\tilde{v}}(t) = F(t,\tilde{v}(t)) - A\tilde{w}(t) + x(t), \tag{1.7a}$$

$$0 = B[\tilde{v}(t) + g(t) + y(t)] \tag{1.7b}$$

with perturbations $x,y : [0,1] \rightarrow \mathbb{R}^{m_1}, y$ differentiable. By using the equivalence between (1.1) and (1.5) and the mean value theorem, it follows that the differences

$$\epsilon_1(t) = \tilde{v}(t) - v(t), \ \epsilon_2(t) = A\tilde{w}(t) - Aw(t)$$

satisfy

$$\dot{\epsilon}_1(t) = (I-C)H(t)\epsilon_1(t) + (I-C)x(t) - C\dot{y}(t),$$

$$\epsilon_2(t) = CH(t)\epsilon_1(t) + Cx(t) + C\dot{y}(t)$$

where

$$H(t) = \int_0^1 F'(t,v(t)+\tau(\tilde{v}(t)-v(t)))d\tau$$

and $F'(t,v)$ stands for the Jacobian matrix $D_vF(t,v)$. From (1.6) it follows that $\mu[(I-C)H(t)]\leqslant\beta$ and $\|CH(t)\|\leqslant\alpha$, and we get for all $t\in[0,1]$ (see [2], [3], for example)

$$|\epsilon_1(t)|\leqslant e^{\beta t}|\epsilon_1(0)|+\{(\beta t)^{-1}(e^{\beta t}-1)\}t\Delta, \tag{1.8a}$$

$$|\epsilon_2(t)|\leqslant\alpha|\epsilon_1(t)|+\Delta \tag{1.8b}$$

whenever $|(I-C)x(t)-C\dot{y}(t)|\leqslant\Delta$ and $|Cx(t)+C\dot{y}(t)|\leqslant\Delta$ for all t. Here $(\beta t)^{-1}(e^{\beta t}-1)$ should be taken equal to 1 if $\beta t = 0$. Thus we see that $v(t)$ and $Aw(t)$ are stable w.r.t. perturbations x and y for which $|x(t)|$ and $|\dot{y}(t)|$ are bounded. It can also be shown that (1.6) is necessary for the above stability result to hold. Important is that in (1.6) the Lipschitz constants of F,A and B are not involved (only of C and CF). Hence the problem may be arbitrarily stiff.

2. The θ-method

In this section we consider the so-called one-leg version of the θ-method, and we discuss its stability properties. Applied to an ordinary differential equation

$$\dot{u}(t) = G(t,u(t))$$

this method reads

$$u^{n+1} = u^n+hG(t^n+\theta h,(1-\theta)u^n+\theta u^{n+1}).$$

Here, θ is a parameter, $h>0$ is a stepsize and $t^n = nh(n = 0,1,2,...)$. This class of method, contains the implicit midpoint rule ($\theta=\frac{1}{2}$) and the Backward Euler method ($\theta=1$). We assume in the following that $\theta\geqslant\frac{1}{2}$. For $\theta<\frac{1}{2}$ the method is not A-stable and there will be no stability for arbitrarily stiff systems.

Let $t^{n+\theta} = t^n+\theta h, v^{n+\theta} = (1-\theta)v^n+\theta v^{n+1}$ and $w^{n+\theta} = (1-\theta)w^n+\theta w^{n+1}$, where the v^n,w^n denote approximations to $v(t^n),w(t^n)$, respectively. Applying the above method for discretization of (1.1a), we get the following scheme for $n = 0,1,2,...$

$$v^{n+1} = v^n+hF(t^{n+\theta},v^{n+\theta})-hAw^{n+\theta}, \tag{2.1a}$$

$$0 = B(v^{n+1}+g(t^{n+1})). \tag{2.1b}$$

In these relations w^{n+1} does not feature explicitly. From a known v^n we can compute v^{n+1} and $w^{n+\theta}$. The approximation w^{n+1} can then be found by using the recursion

$$w^{n+1} = -\theta^{-1}(1-\theta)w^n+\theta^{-1}w^{n+\theta},\ w^0 \text{ from } (1.5b). \tag{2.2}$$

As with the DAE itself we can eliminate the nonstate variables w, giving us

$$v^{n+1} = v^n+h(I-C)F(t^{n+\theta},v^{n+\theta})-hC\dot{g}^{n+\frac{1}{2}},$$

$$Aw^{n+\theta} = CF(t^{n+\theta},v^{n+\theta})+C\dot{g}^{n+\frac{1}{2}}$$

where

$$\dot{g}^{n+\frac{1}{2}} = h^{-1}(g(t^{n+1})-g(t^n)).$$

These formulas show that application of the θ-method to (1.5) leads to the same process for computing the approximations v^n, only with $\dot{g}^{n+\frac{1}{2}}$ replaced by $\dot{g}(t^{n+\theta})$.

In order to analyze stability of the scheme (2.1) we consider a perturbed version

$$\tilde{v}^{n+1} = \tilde{v}^{n} + hF(t^{n+\theta}, \tilde{v}^{n+\theta}) - hA\tilde{w}^{n+\theta} + h\xi^{n+1}, \tag{2.3a}$$

$$0 = B(\tilde{v}^{n+1} + g(t^{n+1}) + h\eta^{n+1}). \tag{2.3b}$$

Here the perturbations ξ^{n+1}, η^{n+1} may stand for round-off errors or errors caused by not solving exactly the nonlinear equations defined by (2.1), but also local discretization errors may be represented this way. The factors h in front of the perturbations are only for notational convenience. For (2.3) we have an initial value $\tilde{v}^0$, and $B(\tilde{v}^0 + g(0) + h\eta^0) = 0$. Define for all $n \geqslant 0$

$$\epsilon_1^n = \tilde{v}^n - v^n,\ \epsilon_2^n = A\tilde{w}^n - Aw^n,$$

and let $H^n = H(t^{n+\theta}, \tilde{v}^{n+\theta}, v^{n+\theta})$ where

$$H(t,x,y) = \int_0^1 F'(t, y + \tau(x-y))d\tau \quad (\textit{for } t \in \mathbb{R} \text{ and } x,y \in \mathbb{R}^{m_1}). \tag{2.4}$$

By subtraction of (2.1) from (2.3) and application of the mean-value theorem we obtain by some calculations

$$\epsilon_1^{n+1} = \epsilon_1^n + h(I-C)H^n\epsilon_1^{n+\theta} + h(I-C)\xi^{n+1} - hC(\eta^{n+1} - \eta^n), \tag{2.5a}$$

$$\epsilon_2^{n+\theta} = CH^n\epsilon_1^{n+\theta} + C(\xi^{n+1} + \eta^{n+1} - \eta^n) \tag{2.5b}$$

where $\epsilon_j^{n+\theta} = (1-\theta)\epsilon_j^n + \theta\epsilon_j^{n+1} (j = 1,2)$. Relation (2.5a) can be rewriten as

$$\epsilon_1^{n+1} = (I - \theta Z_1^n)^{-1}(I + (1-\theta)Z_1^n)\epsilon_1^n + (I - \theta Z_1^n)^{-1}(h(I-C)\xi^{n+1} - hC(\eta^{n+1} - \eta^n))$$

with $Z_1^n = h(I-C)H^n$. Our assumption (1.6) implies $\mu[Z_1^n] \leqslant h\beta$, and we may conclude that for $\theta \geqslant \frac{1}{2}, \theta h\beta < 1$

$$\|(I - \theta Z_1^n)^{-1}(I + (1-\theta)Z_1^n)\| \leqslant (1 - \theta h\beta)^{-1}(1 + (1-\theta)h\beta),$$

$$\|(I - \theta Z_1^n)^{-1}\| \leqslant (1 - \theta h\beta)^{-1}$$

(see for example [3; Th.2.3.1]). Let Δ be an upper bound for $|\xi^n|, |\eta^n|$ (for all n). By using also $\|CH^n\| \leqslant \alpha, \|C\| \leqslant \gamma$ we then obtain from (2.5) the inequalities

$$|\epsilon_1^{n+1}| \leqslant (1 - \theta h\beta)^{-1}(1 + (1-\theta)h\beta)|\epsilon_1^n| + (1 - \theta h\beta)^{-1}h(1 + 3\gamma)\Delta, \tag{2.6a}$$

$$|\epsilon_2^{n+\theta}| \leqslant \alpha|\epsilon_1^{n+\theta}| + 3\gamma\Delta, \tag{2.6b}$$

and (2.6b) implies

$$|\epsilon_2^{n+1}| \leqslant |\theta^{-1}(1-\theta)||\epsilon_2^n| + \theta^{-1}\alpha|\epsilon_1^{n+\theta}| + 3\gamma\theta^{-1}\Delta. \tag{2.6c}$$

Define $\nu_\theta = 0$ for $\theta > \frac{1}{2}$, and $\nu_{\frac{1}{2}} = 1$. From (2.6a), (2.6c) the following global result follows in a standard way.

THEOREM 2.1. *Consider (2.1), (2.3) with $|\xi^n| \leqslant \Delta, |\eta^n| \leqslant \Delta$ (for all n). Assume $\theta \geqslant \frac{1}{2}$ and (1.6). Then there exist $c, \bar{h} > 0$, only depending on α, β, γ and θ, such that for all $n \geqslant 0$, $0 < h \leqslant \bar{h}$ and $0 \leqslant t_n \leqslant 1$*

$$|\epsilon_1^n| \leqslant e^{\beta t_n(1+ch)}|\epsilon_1^0| + ct_n\Delta,$$

$$|\epsilon_2^n| \leqslant |\theta^{-1}(1-\theta)|^n|\epsilon_2^0| + c(1 + \nu_\theta t_n h^{-1})(|\epsilon_1^0| + \Delta).$$

For $\theta > \frac{1}{2}$ the above theorem shows stability: small initial errors and perturbations cause small global errors. In case $\theta = \frac{1}{2}$ we have $\nu_\theta \neq 0$ and then a factor h^{-1} appears in the upper bound for $|\epsilon_2^n|$. By some authors, for example [8], [10], the midpoint rule is called unstable (the situation with so called index 1 systems is similar). It should be noted however that (i) the instability is weak, (ii) it can be avoided by not using (2.2), and (iii) if the ξ^n, η^n depend smoothly on n, as will be the case if they represent local discretization errors, there may be cancellation of errors which will cause the h^{-1} factor to disappear. We shall discuss this below in some detail.

REMARK 2.2. Under the additional assumption that $(I-C)F(t,Cx)$ satisfies a Lipschitz condition w.r.t. x, it can be shown that, for t_n bounded away from zero, the upper bound for $|\epsilon_1^n|$ in theorem 2.1 still holds if we only have $|h\eta^n| \leqslant \Delta$ (instead of $|\eta^n| \leqslant \Delta$). For the Backward Euler method we then get the same result as obtained in [5,p.86] and [7,p.500] for the BDF methods applied to non stiff problems.

Suppose the vectors ξ^n, η^n represent errors caused by round-off and the nonexact solution of the algebraic equations in (2.1), i.e. (2.1) stands for an ideal process whereas (2.3) is the actual computing process. Let $|\epsilon_j^0| \leqslant c\Delta (j = 1,2)$. If $\theta > \frac{1}{2}$ we see that these errors affect v^n and w^n in a similar way, but for $\theta = \frac{1}{2}$ we get $|\epsilon_1^n| \leqslant c\Delta, |\epsilon_2^n| \leqslant ch^{-1}\Delta$. If $h^{-1}\Delta$ is not small it is not advisable to compute the w^n from (2.1), (2.2) for $\theta = \frac{1}{2}$; the recursion (2.2) then allows a linear error growth leading to the h^{-1} factor (see [8]). For this situation there are some alternatives which avoid the use of (2.2). For example, we can compute the w^n, only at points where output is requested, by using (1.5b). A cheaper way, which requires some more storage, is to compute the w^n by interpolation or extrapolation of the $w^{n+\theta} \approx w(t^n + \theta h)$. The computation of these intermediate vectors $w^{n+\theta}$ in (2.1) is always stable, as can be seen from (2.6b) (and the bound for the $|\epsilon_1^n|$).

If we put in (2.3) $\tilde{v}^n = v(t^n), \tilde{w}^n = w(t^n)$ with $v(t), w(t)$ the exact solution of (1.1) the ξ^n, η^n are local (residual) discretization errors, and theorem 2.1 can be used to prove convegence. With stiff systems the local discretization errors are difficult to estimate, due to the fact that no (moderate) bounds may exist for certain derivatives which arise in Taylor series expansions. This is of course very much problem dependent (see Frank et al. [4] for a detailed discussion). It was proved by KRAAYEVANGER [6] that the θ-method applied to arbitrarily stiff ODEs is convergent with order 1 if $\theta > \frac{1}{2}$ and order 2 if $\theta = \frac{1}{2}$. It follows that in our case the same order of convergence holds for the v^n. If (2.2) is avoided, as indicated before, the same orders can be obtained for the w^n. Here we shortly discuss the process (2.1), (2.2) and we assume that all arising derivatives can be bounded properly. This will hold if F satisfies a Lipschitz condition and is sufficiently smooth. Taylor series expansion then shows

$$\eta^n = 0,\ \xi^n = (\tfrac{1}{2} - \theta)h\ddot{v}(t^n) + \tfrac{1}{2}(\tfrac{1}{3} - \theta^2)h^2\dddot{v}(t^n) + ...$$

so that theorem 2.1 can be applied with

$$\Delta = O(h) \text{ if } \theta > \tfrac{1}{2},\ \ \Delta = O(h^2) \text{ if } \theta = \tfrac{1}{2}.$$

This shows first order convergence for both v^n and w^n in case $\theta > \frac{1}{2}$. If $\theta = \frac{1}{2}$ we get second order convergence for the v^n, but seemingly only first order for the w^n. This last result can be improved. Let $\theta = \frac{1}{2}$ and consider the expression in (2.5b) for $\epsilon_2^{n+\theta}$. The right hand side of (2.5b) will depend smoothly on n. Therefore we have not only

$$\epsilon_2^{n+1} = -\epsilon_2^n + 2\epsilon_2^{n+\theta},\ |\epsilon_2^{n+\theta}| = O(h^2), \tag{2.7}$$

but also

$$\epsilon_2^{n+2} = \epsilon_2^n + 2(\epsilon_2^{n+1+\theta} - \epsilon_2^{n+\theta}), |\epsilon_2^{n+1+\theta} - \epsilon_2^{n+\theta}| = O(h^3). \tag{2.8}$$

Direct use of (2.7) leads to the global bound $|\epsilon_2^n| = O(h)$. From (2.8) however we obtain the second order result $|\epsilon_2^n| = O(h^2)$ (for all n).

3. A PREDICTION-PROJECTION METHOD

When using the θ-method of section 2 we solve at each step new approximations v^{n+1} and $w^{n+\theta}$ simultaneously. This is avoided in the following scheme, with two parameters θ and λ. We first compute a prediction u^{n+1} for v^{n+1},

$$u^{n+1} = v^n + hF(t^{n+\theta}, (1-\theta)v^n + \theta u^{n+1}) - h\lambda A w^n, \tag{3.1a}$$

after which v^{n+1} and w^{n+1} can be solved from the relations

$$v^{n+1} = u^{n+1} - h(1-\theta-\lambda)Aw^n - h\theta A w^{n+1}. \tag{3.1b}$$

$$0 = B(v^{n+1} + g(t^{n+1})). \tag{3.1c}$$

In actual computations we will perform (3.1b), (3.1c) by first solving w^{n+1} from

$$h\theta BAw^{n+1} = B(u^{n+1} + g(t^{n+1})) - h(1-\theta-\lambda)BAw^n. \tag{3.2}$$

Then v^{n+1} can be obtained (explicitly) from (3.1b).

Schemes of the above type were introduced for the Navier-Stokes equations by CHORIN [1] and TEMAM [12]; they considered $\theta = 1$ and $\lambda = 0$. VAN KAN [14] constructed a second order method with $\theta = \frac{1}{2}$ and $\lambda = 1$. In these papers the step (3.1a) was simplified by linearization and splitting techniques. Here we will consider (3.1) with $\theta \geqslant \frac{1}{2}$ and $\lambda \geqslant 0$.

It can be seen from (3.1b), (3.1c) that v^{n+1} is a projection of u^{n+1} onto the plane defined by (3.1c). We assume that this projection, which equals $I - C$, is orthogonal. This holds if

$$A^T = B.$$

In order to compare the stability properties of (3.1) with those of the original method (2.1), we consider the perturbed version

$$\tilde{u}^{n+1} = \tilde{v}^n + hF(t^{n+\theta}, (1-\theta)\tilde{v}^n + \theta\tilde{u}^{n+1}) - h\lambda A\tilde{w}^n + h\zeta^{n+1}, \tag{3.3a}$$

$$\tilde{v}^{n+1} = \tilde{u}^{n+1} - h(1-\theta-\lambda)A\tilde{w}^n - h\theta A\tilde{w}^{n+1} + h\eta^{n+1}, \tag{3.3b}$$

$$0 = B(\tilde{v}^{n+1} + g(t^{n+1}) + h\xi^{n+1}), \tag{3.3c}$$

and we define $\epsilon_1^n = \tilde{v}^n - v^n, \epsilon_2^n = A\tilde{w}^n - Aw^n$. Let ν_θ be as in section 2.

THEOREM 3.1. *Let $\theta \geqslant \frac{1}{2}$, $\lambda \geqslant 0$ and suppose C is an orthogonal projection. Consider (3.1), (3.3) with $|\zeta^n|, |\eta^n|, |\xi^n| \leqslant \Delta$ (for all n) and let $\kappa(h) = |\theta^{-1}(1-\theta)| + 2\alpha\lambda h$. There exist positive constants c and $\bar{h}$, only depending on α, β, λ and θ, such that for all $n \geqslant 0$, $0 < h \leqslant \bar{h}$, $0 \leqslant t_n \leqslant 1$,*

$$|\epsilon_1^n| \leqslant e^{c\beta t_n}|\epsilon_1^0| + ct_n^{\frac{1}{2}}\Delta + c\beta\lambda h(h+\nu_\theta t_n)^{\frac{1}{2}}|\epsilon_2^0|,$$

$$|\epsilon_2^n| \leqslant (\kappa(h)^n + c\beta\lambda(h+\nu_\theta t_n)^{3/2})|\epsilon_2^0| + c(1+\nu_\theta t_n h^{-1})(|\epsilon_1^0| + \Delta).$$

The proof of this theorem will be given at the end of this section. First we discuss some of its features.

Comparision with theorem 2.1 shows that we still have qualitatively the same behaviour. The main difference is that with the prediction-projection method the errors ϵ_1^n are influenced by ϵ_2^0, and, as a consequence of this, the initial error ϵ_2^0 remains significant for the ϵ_2^n (for large n), even if $\theta > \frac{1}{2}$. As before there is for $\theta = \frac{1}{2}$ a weak instability giving rise to the h^{-1} factor in the estimate for the ϵ_2^n. In the same way as in section 2 this weak instability can be avoided. (The fact that also with method (3.1) the intermediate vectors $w^{n+\theta}$ are stable can be seen from formula (3.9) together with the above estimates).

For the method (3.1) it is somewhat surprising that the weak instability for the w^n in case $\theta = \frac{1}{2}$

does not influence the v^n too much. After all, unlike method (2.1), the w^n are used explicitly in (3.1). It is the extra assumption that C is orthogonal which is responsible for this (see remark 3.2).

There is a second surprising feature in theorem 3.1. If $\beta = 0$ we have

$$|\epsilon_1^n| \leqslant |\epsilon_1^0| + c\Delta,$$

and thus the accuracy of the v^n is then not influenced by the choice of λ nor by the error in w^0. The reason for this is that there is a certain form of decoupling if $\beta = 0$; in fact it can be shown that the stability assumption (1.6) with $\beta = 0$ implies that $(I-C)F(t,x)$ only depends on $(I-C)x$, not on Cx, and this implies that the v^n computed from (3.1) are identical to those from the method (2.1). (We shall not prove this result since this situation, (1.6) with $\beta = 0$, seems unrealistic).

Theorem 3.1 can be used to prove convergence for method (3.1) by taking $\tilde{v}^n = v(t^n)$, $\tilde{w}^n = w(t^n)$. We are free to choose the vectors $\tilde{u}^n$ in (3.3) in a convenient way. For example, we can take $\tilde{u}^{n+1}$ to be equal to the solution at $t = t^{n+1}$ of

$$\dot{u}(t) = F(t,u(t)) - \lambda A w(t^n) \ (t \geqslant t^n), u(t^n) = v(t^n).$$

Assuming certain derivatives to be bounded (see section 2) it follows by a Taylor series expansion and some calculations that

$$\zeta^{n+1} = (\tfrac{1}{2}-\theta)O(h) + O(h^2), \eta^{n+1} = (1-\lambda)O(h) + (\tfrac{1}{2}-\theta)O(h) + O(h^2), \xi^{n+1} = 0.$$

Thus we can apply theorem 3.1 with $\Delta = O(h^2)$ if $\theta = \frac{1}{2}, \lambda = 1$, and with $\Delta = O(h)$ if $\theta \neq \frac{1}{2}$ or $\lambda \neq 1$, showing convergence for the v^n with order 2 of $\theta = 1/2, \lambda = 1$ and order 1 otherwise. In a similar way as in section 2 it can be shown that this also holds for the w^n (for $\theta = \frac{1}{2}$ the situation is here somewhat more complicated; one can use formula (3.7b)).

Proof of theorem 3.1. Let $\epsilon_0^{n+1} = \tilde{u}^{n+1} - u^{n+1}$ and $Z^n = hH(t^{n+\theta}, (1-\theta)\tilde{v}^n + \theta\tilde{u}^{n+1}, (1-\theta)v^n + \theta u^{n+1})$ (see (2.4)). From (3.1) and (3.3) we obtain by subtraction

$$\epsilon_0^{n+1} = \epsilon_1^n + Z^n[(1-\theta)\epsilon_1^n + \theta\epsilon_0^{n+1}] - h\lambda\epsilon_2^n + h\zeta^{n+1}, \tag{3.4a}$$

$$\epsilon_1^{n+1} = \epsilon_0^{n+1} - h(1-\theta-\lambda)\epsilon_2^n - h\theta\epsilon_2^{n+1} + h\eta^{n+1} \tag{3.4b}$$

$$0 = B[\epsilon_1^{n+1} + h\xi^{n+1}]. \tag{3.4c}$$

By eliminating ϵ_0^{n+1} from (3.4a), (3.4b) we obtain

$$\begin{aligned} &\epsilon_1^{n+1} - \epsilon_1^n + h\epsilon_2^{n+\theta} - h\eta^{n+1} - h\zeta^{n+1} = \\ &= Z^n[\epsilon_1^{n+\theta} + h\theta\epsilon_2^{n+\theta} - h\theta\lambda\epsilon_2^n - h\theta\eta^{n+1}] \end{aligned} \tag{3.5a}$$

Further we have for all n

$$(I-C)\epsilon_2^n = 0, \ C\epsilon_1^n = -hC\xi^n \tag{3.5b}$$

The recursion formed by (3.5) becomes a bit more simple to handle by introducing

$$e_1^n = \epsilon_1^n + h\xi^n, \ e_2^n = h\theta\epsilon_2^n$$

and

$$d^{n+1} = (I-\theta Z^n)^{-1}[(\eta^{n+1} + \zeta^{n+1} + \xi^{n+1} - \xi^n) - \theta Z^n(\eta^{n+1} + \xi^{n+1} + \theta^{-1}(1-\theta)\xi^n)].$$

Then

$$e_1^{n+1} - e_1^n + \theta^{-1}e_2^{n+\theta} - hd^{n+1} = Z^n[e_1^{n+\theta} + e_2^{n+\theta} - \lambda e_2^n - \theta h d^{n+1}], \tag{3.6a}$$

$$(I-C)e_2^n = 0 \text{ and } Ce_1^n = 0. \tag{3.6b}$$

Let

$$d_1^{n+1} = (I-C)d^{n+1}, d_2^{n+1} = Cd^{n+1}, Z_1^n = (I-C)Z^n \text{ and } Z_2^n = CZ^n.$$

From (3.6) it follows that

$$e_1^{n+1} - e_1^n - hd_1^{n+1} = Z_1^n[e_1^{n+\theta} + e_2^{n+\theta} - \lambda e_2^n - \theta h d^{n+1}], \tag{3.7a}$$

$$e_2^{n+\theta} - \theta h d_2^{n+1} = \theta Z_2^n[e_1^{n+\theta} + e_2^{n+\theta} - \lambda e_2^n - \theta h d^{n+1}]. \tag{3.7b}$$

In the following we shall use c to denote a positive constant depending on α,β,λ and θ, not necessarily always with the same value. Further it will be tacitly assumed that the stepsize h is bounded from above such that arising terms like $(1-ch)^{-1}$ can be bounded by a constant for all possible h. From (1.6) it follows that $\mu[Z^n]\leqslant h(\alpha+\beta)$. Application of theorem 2.3.1 in [3] shows that

$$|d^n| \leqslant c\Delta \text{ (for all } n). \tag{3.8}$$

Now consider (3.7b). Since $\|Z_2^n\|\leqslant\alpha h$ we have

$$|e_2^{n+\theta} - \theta h d_2^{n+1}| \leqslant \theta\alpha h\{|e_1^{n+\theta} - \theta h d_1^{n+1}| + |e_2^{n+\theta} - \theta h d_2^{n+1}| + |\lambda e_2^n|\}.$$

Hence, assuming $(1-\theta\alpha\bar{h})^{-1}\leqslant 2$,

$$|e_2^{n+\theta} - \theta h d_2^{n+1}| \leqslant 2\theta\alpha h\{|e_1^{n+\theta} - \theta h d_1^{n+1}| + |\lambda e_2^n|\}, \tag{3.9}$$

$$|e_2^{n+1}| \leqslant \kappa(h)|e_2^n| + ch\{|e_1^{n+1}| + |e_1^n| + \Delta\}. \tag{3.10}$$

Next we consider (3.7a). We have $\mu[Z_1^n]\leqslant h\beta$. Further, since C is an orothogonal projection, $(x,y) = 0$ whenever $x,y\in\mathbb{R}^{m_1}, (I-C)x = x$ and $Cy = y$. It follows that

$$(e_1^{n+1} - e_1^n - hd_1^{n+1}, e_1^{n+\theta} - \theta h d_1^{n+1}) \leqslant h\beta|e_1^{n+\theta} + e_2^{n+\theta} - \lambda e_2^n - \theta h d^{n+1}|^2 =$$
$$= h\beta\{|e_1^{n+\theta} - \theta h d_1^{n+1}|^2 + |e_2^{n+\theta} - \lambda e_2^n - \theta h d_2^{n+1}|^2\}.$$

For any two vectors $x,y\in\mathbb{R}^m$ and $\theta\geqslant\frac{1}{2}$ we have the inequality

$$(x-y, \theta x + (1-\theta)y) \geqslant \frac{1}{2}|x|^2 - \frac{1}{2}|y|^2.$$

This follows by evaluating the left-hand side and using $(x,y)\leqslant\frac{1}{2}|x|^2+\frac{1}{2}|y|^2$. Application of the inequality with $x = e_1^{n+1} - hd_1^{n+1}$ and $y = e_1^n$ shows that

$$|e_1^{n+1} - hd_1^{n+1}|^2 - |e_1^n|^2 \leqslant 2\beta h\{|e_1^{n+\theta} - \theta h d_1^{n+1}|^2 + |e_2^{n+\theta} - \lambda e_2^n - \theta h d_2^{n+1}|^2\}.$$

Hence

$$|e_1^{n+1} - hd_1^{n+1}|^2 - |e_1^n|^2 \leqslant 4\beta h\{\theta^2|e_1^{n+1} - hd_1^{n+1}|^2 + \tag{3.11}$$
$$+ (1-\theta)^2|e_1^n|^2 + |e_2^{n+\theta} - \theta h d_2^{n+1}|^2 + |\lambda e_2^n|^2\}.$$

In case $\beta = 0$ the statement of the theorem easily follows from (3.10) and (3.11). Assume in the following $\beta>0$. We have

$$|e_1^{n+1} - hd_1^{n+1}|^2 \geqslant |e_1^{n+1}|^2 - 2h|e_1^{n+1}||d_1^{n+1}| + h^2|d_1^{n+1}|^2 \geqslant$$
$$\geqslant (1-h)|e_1^{n+1}|^2 - h(1-h)|d_1^{n+1}|^2.$$

Combining this with (3.8), (3.9) and (3.11) we obtain

$$|e_1^{n+1}|^2 \leqslant (1+c\beta h)|e_1^n|^2 + c\beta h|\lambda e_2^n|^2 + ch\Delta^2. \tag{3.12}$$

We shall use (3.10) and (3.12) to get the upper bounds for $|e_1^n|$ and $|e_2^n|$ in terms of the data. Without loss of generality it may be assumed that the sequence $\{|e_1^n|\}$ is nondecreasing. From (3.10) we then easily obtain

$$|e_2^n| \leqslant \kappa(h)^n|e_2^0| + c(h+\nu_\theta t_n)(|e_1^n| + \Delta). \tag{3.13}$$

Insertion of this inequality into (3.12) leads to

$$|e_1^{n+1}|^2 \leqslant (1+c\beta h)|e_1^n|^2 + c\beta h\kappa(h)^{2n}|\lambda e_2^0|^2 + ch\Delta^2,$$

and we obtain the global result

$$|e_1^n|^2 \leqslant e^{c\beta t_n}|e_1^0|^2 + ctn\Delta^2 + c\beta(h+\nu_\theta t_n)|\lambda e_2^0|^2,$$

under the assumption that $\bar{h}$ is such that $\kappa(\bar{h})<1$ if $\theta>\frac{1}{2}$. Taking square roots on both sides we thus get

$$|e_1^n| \leqslant e^{c\beta t_n}|e_1^0| + ct_n^{\frac{1}{2}}\Delta + c\beta\lambda(h+\nu_\theta t_n)^{\frac{1}{2}}|e_2^0|. \tag{3.14}$$

By combining (3.13) and (3.14) we further obtain

$$|e_2^n| \leqslant (\kappa(h)^n + c\beta\lambda(h+\nu_\theta t_n)^{3/2})|e_2^0| + c(h+\nu_\theta t_n)(|e_1^0|+\Delta). \tag{3.15}$$

The inequalities of the theorem now follow easily. □

REMARK 3.2. In this section we have made the extra assumption that C and $I-C$ are orthogonal projections. In case this does not hold method (3.1) can still be applied, although the method is less natural then (v^{n+1} is then no longer the orthogonal projection of u^{n+1} onto the plane defined by the algebraic constraints). It can be shown that we have stability for nonorthogonal projections provided that either $\theta>\frac{1}{2}$ or $\lambda = 0$. However, for the interesting case $\theta = \frac{1}{2}$, $\lambda = 1$ the weak instability for the w^n may then influence the v^n. This can be seen by some tedious calculations with a linear test-problem and $m_1 = 2, m_2 = 1$. We omit the proof of these statements since nonorthogonal projections seem of minor importance.

4. CONCLUDING REMARKS

The restriction to the θ-methods as ODE method in the foregoing was only imposed to be able to derive stability results. The prediction-projection idea can also be used for higher order multistep methods.

For the derivation of the stability results for the θ-method and its modification (3.1) we have used assumption (1.6), which guarantees stability in the Euclidian norm of the DAE system itself in a rather general sense. Under this assumption both methods appear to be stable (with some restrictions if $\theta=\frac{1}{2}$). This situation would change with other assumptions. Suppose, for instance, that $g\equiv 0$, so that the exact solution v of (1.1) lies in the plane defined by the equation $Bu = 0$. The function $F(t,\cdot)$ may only be defined in a reasonable way near this plane. If the DAE (1.1) is only stable as long as v remains very close to this plane then the θ-method (2.1) may still be stable, but with method (3.1) difficulties may be expected, due to the fact that the prediction u^{n+1} leaves the plane. Numerical experiments on the Navier-Stokes equations by VAN KAN [14] and TEN THIJE BOONKKAMP [13] with prediction-projection schemes similar to (3.1) (with different ODE methods) revealed no stability problems, so that assumption (1.6) seems more realistic for this equation than the situation discussed above.

Finally we note that in [14] stability has been proved for method (3.1) with $\theta=1/2, \lambda=1$ under the assumption that $A^T=B, F$ is linear and time independent, and

$$\mu[F] \leqslant 0\,. \tag{4.1}$$

We have not followed this approach here. Condition (4.1) does not guarantee stability of (1.1) (under perturbations). One can add the assumption $\|CF\|\leqslant\alpha$. This is necessary for small errors in $v^0\approx v(0)$ to cause only small errors too in w^0 (cf. (1.5b)). However, then we have $\mu[(I-C)F]\leqslant\alpha$, which implies that our assumption (1.6) holds with $\alpha=\beta$. In this sense (1.6) is more general than (4.1).

References

1. A.J. Chorin, *On the convergence of discrete approximations to the Navier-Stokes equations.* Math. Comp. 23 (1969), 341-353.
2. G. Dahlquist, *Error analysis for a class of methods of stiff non-linear initial value problems.* Lecture Notes in Mathematics 506, Springer Verlag, Berlin, 1976.
3. K. Dekker and J.G. Verwer, *Stability of Runge-Kutta methods for stiff nonlinear differential equations.* North-Holland, Amsterdam, 1984.
4. R. Frank, J. Schneid and C.W. Ueberhuber, *The concept of B-convergence.* SIAM J. Numer. Anal. 18 (1981), 753-780.
5. C.W. Gear, B. Leimkuhler and G.K. Gupta, *Automatic integration of Euler-Lagrange equations with constraints.* J. Comp. Appl. Math 12, 13 (1985), 77-90.
6. H.F.B.M. Kraaijevanger, *B-convergence of the implicit midpoint rule and the trapezoidal rule.* BIT 25 (1985), 652-666.
7. P. Lötstedt and L.R. Petzold, *Numerical solution of nonlinear differential equations with algebraic constraints I: convergence results for BDF methods.* Math. Comp. 46 (1986), 491-516.
8. R. März, *On difference and shooting methods for boundary value problems in differential-algebraic equations.* ZAMM 64 (1984), 463-473.
9. R. März, *On initial value problems in differential-algebraic equations and their numerical treatment.* Computing 35 (1985), 13-37.
10. L.R. Petzold, *Order results for implicit Runge-Kutta methods applied to differential-algebraic systems.* SIAM J. Numer. Anal. 23 (1986), 837-852.
11. L.R. Petzold and P. Lötstedt, *Numerical solution of nonlinear differential equations with algebraic constraints,* part II: *practical implications.* SIAM J. Sci. Stat. Comp. 7 (1986), 720-734.
12. R. Témam, *Sur l'approximation de la solution des équations de Navier-Stokes par la méthode des pas fractionnaires* (II). Arch. Rat. Mech. Anal. 33 (1969), 377-385.
13. J.H.M. ten Thije Boonkkamp, *The odd-even hopscotch pressure correction scheme for the incompressible Navier-Stokes equations.* Report NM-R8615, Centre for Math. and Comp. Sc., Amsterdam, 1986.
14. J. van Kan, *A second order pressure-correction method for viscous incompressible flow.* SIAM J. Sci. Stat. Comp. (1986), 870-891.