stichting
mathematisch
centrum

MC

AFDELING NUMERIEKE WISKUNDE (DEPARTMENT OF NUMERICAL MATHEMATICS)

NW 100/81 MAART

J.G. VERWER

INSTRUCTIVE EXPERIMENTS WITH SOME RUNGE-KUTTA-ROSENBROCK METHODS

Preprint

kruislaan 413 1098 SJ amsterdam

Printed at the Mathematical Centre, 413 Kruislaan, Amsterdam.

The Mathematical Centre, founded the 11-th of February 1946, is a non-profit institution aiming at the promotion of pure mathematics and its applications. It is sponsored by the Netherlands Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).

1980 Mathematics subject classification: 65L05

ACM-Computing Reviews-category: 5.17

Instructive experiments with some Runge-Kutta-Rosenbrock methods*)

by

J.G. Verwer

ABSTRACT

The paper deals with certain boundedness properties of Runge-Kutta-Rosenbrock methods when applied to nonlinear stiff systems. It reports some instructive examples and numerical experiments performed with a number of simple 2-stage schemes and the Rosenbrock code ROW4A. Attention is paid to the conversion of non-autonomous problems to the autonomous form. An important conclusion is that this conversion may lead to a significant loss in accuracy.

KEY WORDS & PHRASES: *Numerical analysis, Numerical integration, Rosenbrock methods, Nonlinear stiff equations*

---

*) This report will be submitted for publication elsewhere.

1. INTRODUCTION

A substantial part of the literature on numerical methods for stiff systems of ordinary differential equations deals with Runge-Kutta-Rosenbrock methods. For the *non-autonomous* initial value problem

$$\dot{X} = F(t,X), \quad X(t_0) = X_0, \tag{1.1}$$

the *original* m-stage Rosenbrock method (see [7]) is very similar to the Runge-Kutta type integration formula

$$\begin{aligned} X_n^{(0)} &= X_n, \\ K_n^{(j)} &= [I - \gamma_j \tau J_n^{(j)}]^{-1} F(t_n^{(j)}, X_n^{(j)}), \quad \gamma_j > 0, \ j = 0(1)m-1, \\ X_n^{(j)} &= X_n + \tau \sum_{\ell=0}^{j-1} \lambda_{j,\ell} K_n^{(\ell)}, \quad j = 1(1)m, \\ X_{n+1} &= X_n^{(m)}, \quad n = 0,1,\ldots . \end{aligned} \tag{1.2}$$

X_n denotes the approximation at time $t = t_n$ and $\tau > 0$ denotes the stepsize; $t_n^{(j)} = t_n + \nu_j \tau$, where, normally, $0 \le \nu_j \le 1$. Further

$$\begin{aligned} J_n^{(j)} &= J(\hat{t}_n^{(j)}, \hat{X}_n^{(j)}), \quad J(t,X) = \partial F(t,X)/\partial X, \\ \hat{t}_n^{(j)} &= \sum_{\ell=0}^{j} \alpha_{j,\ell} t_n^{(\ell)}, \quad \hat{X}_n^{(j)} = \sum_{\ell=0}^{j} \alpha_{j,\ell} X_n^{(\ell)}, \end{aligned} \tag{1.3}$$

where the parameters $\alpha_{j,\ell}$ denote real scalars. Note that each stage involves an F(t,X)-evaluation, a solution of a system of linear algebraic equations, and, possibly, a J(t,X)-evaluation.

Up to now the literature on Rosenbrock type schemes mainly deals with the development of new schemes and, in particular, with the analysis of the appearing rational stability functions. In fact, it is now well-known that there do exist A-stable, or L-stable, Rosenbrock type schemes of high order of consistency. It is less known however that such a scheme, which according to the Dahlquist-Henrici theory ought to be judged as being reliable, may

behave real bad when applied to certain non-linear problem classes. Or, when we are given 2 schemes of the same order of consistency and having the same stability function, we may encounter large differences in their performance when applied to these non-linear problem classes.

The present paper deals with these phenomena. We discuss a number of instructive examples and numerical experiments, most of which are based on results presented in a previous paper [12]. In that paper the author investigated, following ideas put forward by Stetter [10] and van Veldhuizen [11], a so-called *uniform boundedness property* of method (1.2) for 2 model classes which are directly relevant to non-linear stiff problems. This boundedness property plays a key role in the examples and experiments we are going to discuss.

In short, the contents of the paper are as follows. In section 2 we shortly discuss the boundedness property we are concentrating on. Sections 3 and 4 review the model classes we investigated in [12]. In these sections we also discuss numerical examples. Section 5 deals with the conversion to the autonomous form which in the greater part of the literature is used when a genuine non-autonomous problem is met. An important conclusion of section 5 is that this conversion to the autonomous form may lead to a significant loss in accuracy, and even to instability. In section 5 we also report an experiment with the Rosenbrock code ROW4A. Here our aim is to illustrate how bad boundedness properties show up in practice when using an automatic code.

2. THE PROPERTY OF ε-BOUNDEDNESS

In the analysis of numerical methods for stiff problems the study of *model-equations* have proved to be fruitful. For example, the simple well-known scalar model

$$\dot{x} = \delta x, \quad \delta \in \mathbb{C}, \quad \mathrm{Re}(\delta) < 0, \tag{2.1}$$

provides indispensable information on the absolute stability of integration methods for ordinary differential systems. For constant coefficient linear systems this scalar model already yields enough insight. For *non-linear*

stiff systems however, this model has appeared to be too simple and there is a need for additional research on more refined models. Such a model (cf. [10,11]) should permit the simultaneous *occurrence of smooth and transient solution components*, and, in this connection, its Jacobian matrix should have a *time-dependent eigensystem*. Further it should be possible to consider a limit process by which one can introduce *arbitrarily high stiffness*. Finally, the *occurrence of non-linear terms* in the model could help us to increase our insight.

It is our purpose to support these views for Rosenbrock type methods by means of some instructive examples and numerical experiments. Most of these will be based on theoretical results presented in [12]. There we investigated a so-called property of uniform boundedness for method (1.2) when applied to 2 model classes having the characteristics just mentioned. We shall now first describe the kind of boundedness we think of. Let

$$\dot{X} = F(t,X,\varepsilon), \quad \varepsilon \in (0,\varepsilon_0], \quad \varepsilon_0 \text{ constant}, \tag{2.2}$$

represent some class of model equations we have in mind, where

a) $t \in [t_0,T]$, t_0 and T finite and constant, $X(t_0) = X_0 = X_0(\varepsilon)$.

b) All problems in this class possess a unique bounded solution $X = X(t,\varepsilon)$ on $[t_0,T] \times (0,\varepsilon_0]$, i.e., we suppose the existence of a constant K such that

$$\sup_{\varepsilon\in(0,\varepsilon_0]} \sup_{t\in[t_0,T]} \|X(t,\varepsilon)\| \leq K.$$

c) The stiffness ratio tends to infinity if $\varepsilon \to 0$ ($1/\varepsilon$ factors).

Note that the initial vector X_0 may depend on the stiffness parameter ε. This case may be relevant in case we have non-linearities in X. In what follows it is convenient to represent scheme (1.2) in the operator form

$$X_n^{(0)} = X_n,$$

$$X_n^{(j)} = \Phi^{(j)}(\{t_n^{(\ell)}, X_n^{(\ell)}\}_{\ell<j}, \tau, \varepsilon; F), \quad j = 1(1)m, \tag{2.3}$$

$$X_{n+1} = X_n^{(m)}$$

DEFINITION 2.1. Suppose we are given a method of type (2.3) and a class of stiff problems satisfying properties (2.2a-c). We then call this method ε-bounded on this class if for all its problems the following statement holds: for any point (t,X) in the region of definition of F, where $X = X(\varepsilon)$ is bounded in $\varepsilon \in (0,\varepsilon_0]$, a constant τ^* exists such that

$$\Phi^{(j)}(\{t^{(\ell)}, X^{(\ell)}\}_{\ell<j}, \tau, \varepsilon; F) = O(1), \quad \varepsilon \to 0, \quad j = 1(1)m, \tag{2.4}$$

for all $\tau \in (0,\tau^*]$, τ^* being independent of ε. □

For clarity we wish to make 2 comments on this definition. Firstly, in relation (2.4) we confine ourselves to fixed τ-values, i.e., the constant implied may depend on τ (cf. [10], p. 192). In view of property (2.2b), our goal is to select methods which are able to produce a finite sequence of approximations over the interval $[t_0,T]$ being bounded in $\varepsilon \in (0,\varepsilon_0]$. If, for a given problem, none finite sequence will remain bounded if $\varepsilon \to 0$, we may expect large discretization errors in a non-limit situation.

Our second comment concerns the additional boundedness requirement for $j < m$. We prefer to define ε-boundedness in this way as it facilitates the analysis (see [12]) and, of course, it is also obvious to ask for boundedness of $\Phi^{(j)}$, $j < m$, if $\Phi^{(m)}$ is required to be bounded (in general $\Phi^{(m)}$ depends in a non-linear way on $\Phi^{(j)}$, $j < m$).

3. MODEL CLASS 1

In order to obtain concrete results on ε-boundedness one has to select appropriate model classes. In [12] we investigated 2 such classes. The first of these is reviewed in section 3.1. In section 3.2 we present a specific example to be used in section 3.3 for a numerical illustration.

3.1. A class of non-linear model equations

The class is described by 2 coupled singularly perturbed differential systems of the form (see also [3])

$$\begin{aligned} \dot{x} &= f(t,x,y,\varepsilon) + \varepsilon^{-1}A(t)y, \quad x(0) = x_0, \\ \dot{y} &= g(t,x,y,\varepsilon) + \varepsilon^{-1}\mu(t)By, \quad y(0) = y_0. \end{aligned} \tag{3.1}$$

We consider (3.1) on the interval $[0,T]$ and, until further notice, x_0, y_0 are assumed to be independent of ε. The right hand side functions are supposed to be sufficiently differentiable. The vector functions f and g are allowed to be non-linear and, in particular, they are supposed to be bounded in ε as $\varepsilon \to 0$. Further, $f: [0,T] \times \mathbb{R}^{s_1} \times \mathbb{R}^{s_2} \times (0,\varepsilon_0] \to \mathbb{R}^{s_1}$ and $g: [0,T] \times \mathbb{R}^{s_1} \times \mathbb{R}^{s_2} \times (0,\varepsilon_0] \to \mathbb{R}^{s_2}$, where $s_1, s_2 \geq 1$. A is a t-dependent (s_1,s_2)-matrix and μ is a scalar function which is strictly positive, i.e., $\mu(t) \geq \tilde{\mu} > 0$ for all $t \in [0,T]$. Finally, B is a constant (s_2,s_2)-matrix whose spectrum $\Lambda(B)$ lies in the negative half plane $\mathbb{C}^- = \{z \mid \mathrm{Re}(z) < 0\}$. It is not difficult to prove the following result [12]:

THEOREM 3.1. *Let* $\alpha = \max\{\mathrm{Re}(\lambda) : \lambda \in \Lambda(B)\} < 0$. *Then, for all* $t \in (0,T]$ *and* $\varepsilon \in (0,\varepsilon_0]$, *the solution functions* $x(t,\varepsilon)$ *and* $y(t,\varepsilon)$ *of problem* (3.1) *satisfy*

$$\begin{aligned} &\|x(t,\varepsilon)\| \leq K_0, \quad \|\dot{x}(t,\varepsilon)\| \leq K_1[\varepsilon^{-1}\exp(\tfrac{1}{2}\alpha\tilde{\mu}\varepsilon^{-1}t) + 1], \\ &\|y(t,\varepsilon)\| \leq \tilde{K}_0[\exp(\tfrac{1}{2}\alpha\tilde{\mu}\varepsilon^{-1}t) + \varepsilon], \\ &\|\dot{y}(t,\varepsilon)\| \leq \tilde{K}_1[\varepsilon^{-1}\exp(\tfrac{1}{2}\alpha\tilde{\mu}\varepsilon^{-1}t) + 1], \end{aligned} \tag{3.2}$$

K_0, $\tilde{K}_0$, K_1 *and* $\tilde{K}_1$ *being positive constants independent of* t *and* ε. □

These inequalities reveal that we can write

$$x(t,\varepsilon) = O(1), \quad y(t,\varepsilon) = O(\varepsilon), \quad \varepsilon \to 0, \quad t \in (0,T]. \tag{3.3}$$

Normally the x-solution shall consist of a rapidly decaying transient

component and a smooth one which determines $x(t,\varepsilon)$ everywhere outside the transient phase. The transient behaviour of $x(t,\varepsilon)$ is completely determined by the transient of the y-solution. Further, to a large extent the magnitude of the smooth component is independent of the stiffness parameter ε. For the y-solution the situation is somewhat different. Typically, it contains a transient component and a smooth one which is $O(\varepsilon)$ for all $t \in (0,T]$. Hence in a practical situation it will be smooth x-solution in which we are mostly interested, ε being so small that the transients can be neglected and that the smooth y-solution is of less practical interest. It shall be clear now that a suitable integration method for (3.1) should generate approximations to the smooth solutions which show a similar behaviour in ε. In particular, the method should be capable to generate such approximations with some step-size τ being independent of ε, i.e. $X_n^{(j)} = [x_n^{(j)}, y_n^{(j)}]^T$, $j = 1(1)m$, should satisfy

$$(3.4) \qquad x_n^{(j)} = O(1), \quad y_n^{(j)} = O(\varepsilon) \text{ as } \varepsilon \to 0, \; n = 1(1)T/\tau.$$

DEFINITION 3.1. Suppose we are given a method (2.3) which is ε-bounded on a class of problems of type (3.1). We then call this method ε-accurate on this class, if in relations (2.4) for all y-components of $\Phi^{(j)}$ an $O(\varepsilon)$ behaviour appears. □

Clearly, if a method is ε-accurate it can be used to generate finite approximation sequences satisfying (3.4). The next theorem summarizes the main results we obtained for method (1.2) when applied to class (3.1) [12]:

THEOREM 3.2. (i) *Any Rosenbrock method* (1.2) *is* ε*-bounded on the* 2 *classes of problems* (3.1) *for which, respectively,* $A = 0$ *and* A,μ *are constant.*
(ii) *Any Rosenbrock method* (1.2) *is* ε*-bounded on the whole class* (3.1), *if at each stage* $J(t,X)$ *is evaluated at the special point* $(t,X) = (t^{(j)},X^{(j)})$.
(iii) *Any Rosenbrock method* (1.2) *is* ε*-accurate on the whole class* (3.1), *iff the stability function* $R^{(m)}(z)$, *as well as all internal stability functions* $R^{(j)}(z)$, $j < m$, *do have a zero at infinity.*
(iv) *Any Rosenbrock method* (1.2) *evaluating* $J(t,x)$ *once per step, is* ε*-bounded on the whole class* (3.1), *iff* $R^{(j)}(\infty) = 0$ *for* $j < m$.

(v) *Consider class* (3.1). *Let the point* $X = (x,y)$ *occurring in Definition* 2.1 *be such that* $x = O(1)$, $y = O(\varepsilon)$. *Then any Rosenbrock method* (1.2) *is* ε*-accurate on the whole class* (3.1). □

REMARK 3.1. As shown in [12], ε-boundedness of (1.2) with respect to (3.1), is determined by the boundedness, in $\varepsilon \in (0,\varepsilon_0]$, of

$$\varepsilon^{-1}A(t)y + \gamma\tau\varepsilon^{-2}A(\hat{t})[I-\gamma\tau\varepsilon^{-1}\mu(\hat{t})B]^{-1}\mu(t)By, \quad t \neq \hat{t}. \tag{3.5}$$

We shall use this rule to select an appropriate example model for the experiments. It is needed because for a specific example the conditions of Theorem 3.2 may happen to be too strong. □

3.2. A non-linear test example

We consider the system

$$\begin{aligned} \dot{x}_1 &= a_1(x_1+x_2+y-1)^k + \varepsilon^{-1}\mu_1(t)y, \\ \dot{x}_2 &= a_2(x_1+x_2+y-1)^k + \varepsilon^{-1}\mu_2(t)y, \\ \dot{y} &= a_3(x_1+x_2+y-1)^k - \varepsilon^{-1}\mu(t)y. \end{aligned} \tag{3.6}$$

Here $t_0 \le t \le T$ and $\varepsilon \in (0,\varepsilon_0]$, $x_1(t),x_2(t),y(t)$ are scalar, a_i and $k \ge 1$ are constant, and $\mu = \mu_1 + \mu_2$. The sum $s = x_1+x_2+y$ satisfies

$$\dot{s} = a(s-1)^k, \quad a = a_1+a_2+a_3, \tag{3.7}$$

so that

$$s(t) = 1 + [a(1-k)t + C]^{\frac{1}{1-k}}, \quad C = (s(t_0)-1)^{1-k} - a(1-k)t_0. \tag{3.8}$$

If $s(0) = s_0 \neq 1$, equation (3.6) thus possesses a unique solution being bounded on any finite interval $[0,T]$, uniformly in $\varepsilon \in (0,\varepsilon_0]$. Furthermore, this solution satisfies the inequalities in Theorem 3.1.

Elaborating expression (3.5) for system (3.6) yields

$$(3.9)\qquad \varepsilon^{-1}\mu_i(t)y\left\{\frac{\mu_i(t)+\gamma\tau\varepsilon^{-1}[\mu_i(t)\mu(\hat{t})-\mu_i(\hat{t})\mu(t)]}{\mu_i(t)[1+\gamma\tau\varepsilon^{-1}\mu(\hat{t})]}\right\},\ i=1,2.$$

If $\mu_i(t)\mu(\hat{t}) \neq \mu_i(\hat{t})\mu(t)$, this expression is not bounded in $\varepsilon \in (0,\varepsilon_0]$, i.e., the conditions of Theorem 3.2 apply to the specific example (3.6). If $\mu_i(t)\mu(\hat{t}) = \mu_i(\hat{t})\mu(t)$ for all $t,\hat{t} \in [0,T]$, any Rosenbrock method (1.2) is able to generate finite approximation sequences being bounded in $\varepsilon \in (0,\varepsilon_0]$.

The eigenvalues of the Jacobian $\partial F(t,X,\varepsilon)/\partial X$, evaluated on the exact solution, are given by

$$(3.10)\qquad \delta_1 = 0,\ \delta_2 = a\theta^{-1}(t),\ \delta_3 = -\varepsilon^{-1}\mu(t),$$

where $\theta(t) = k^{-1}[at(1-k)+C]$. In the following we therefore take $C > 0$ and $a < 0$, so that $\delta_2 < 0$. Note that δ_2 does not depend on ε.

Obviously, much freedom is left in choosing the various defining parameters in (3.6). We put ($\mu = \mu_1 + \mu_2$)

$$(3.11)\qquad k = 2,\ a_1 = -0.1,\ a_2 = 1,\ a_3 = -1,\ \mu_1(t) = e^t - t,\ \mu_2 = t.$$

Note that for all $t,\hat{t}$ we have $\mu_i(t)\mu(\hat{t}) \neq \mu_i(\hat{t})\mu(t)$, $i = 1,2$. Further, as $\mu_2(0) = 0$, x_2 has no transient. There remains to choose a range of ε-values and initial values at $t = 0$. The ε-range will be given below at the actual experiments. Here we already define 2 sets of initial values, namely

$$(3.12a)\qquad x_1(0) = 0,\ \ x_2(0) = 1,\ \ y(0) = \frac{1}{4},$$

$$(3.12b)\qquad x_1(0) = \frac{1}{4},\ \ x_2(0) = 1,\ \ y(0) = \varepsilon.$$

The initial values (3.12b) define a smooth solution ($y(0) = \varepsilon$).

3.3. Numerical illustration

The lack of ε-boundedness, or ε-accuracy, manifests itself by unusually large errors and, typically, the smaller ε, the larger the errors. We shall

illustrate this unwanted phenomenon for the problems (3.6,11,12a) and (3.6,11,12b).

For the experiments we selected 4 simple 2-stage formulas (1.2) of order 2. All are L-stable and $R^{(1)}$ and $R^{(2)}$ are given by ($\gamma_0=\gamma_1=\gamma$)

$$(3.13)\qquad R^{(1)}(z)=\frac{1+(\lambda_{10}-\gamma)z}{1-\gamma z}\ ,\quad R^{(2)}(z)=\frac{1+(1-2\gamma)z}{(1-\gamma z)^2}\ ,\quad \gamma = 1-\frac{1}{2}\sqrt{2}\ .$$

Note that the formulas share the stability function $R^{(2)}$. We have $\lambda_{20}=1-\lambda_{21}$, $v_1 = 1/2\lambda_{21}$ and $\lambda_{21} = (\frac{1}{2}-\gamma)/\lambda_{10}$:

(3.14)

formula	λ_{10}	α_{00}	α_{10}	α_{11}	$R^{(1)}(\infty)$	J(t,X)	ε-bounded on (3.1)	ε-accurate on (3.1)
a	$1-2\gamma$	1	1	0	$\frac{3\gamma-1}{\gamma}$	1	no	no
b	γ	1	1	0	0	1	yes	yes
c	$1-2\gamma$	1	0	1	$\frac{3\gamma-1}{\gamma}$	2	yes	no
d	γ	1	0	1	0	2	yes	yes

The choice $\lambda_{10}=\gamma$ implies $R^{(1)}(\infty)=0$. The choice $\lambda_{10}=(3\gamma-1)/\gamma$ is, for our purpose, rather arbitrary. Of importance is that in this case $R^{(1)}(\infty)\neq 0$. The present λ_{10}-value implies $v_1=1$ and $R^{(1)}(\infty)\simeq -0.4$. The $\alpha_{j,\ell}$-values are self-evident. Recall that schemes using more than one J(t,X)-evaluation per step, are usually not recommended.

In the figure below we plotted, for a set of ε-values from the interval $[10^{-7},1]$, the numbers $ac_x = -^{10}\log(\text{max. abs. error of x-components})$ and $ac_y = -^{10}\log(\text{abs. error of y-component})$ for precisely 1 integration step of length 1/20. On purpose we do not give errors measured after a number of steps because we noticed cancellation of x_1-errors and x_2-errors when performing more than 1 step. For our purpose it suffices to consider only 1 step. Recall that problem (3.6,11,12a) exhibits a transient behaviour, whereas the solution of (3.6,11,12b) is smooth due to the initial value $y(0)=\varepsilon$.

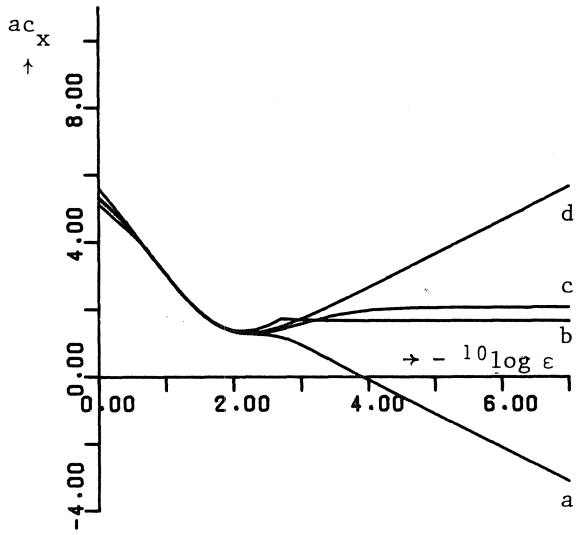

Fig. 3.1 Initial x-error, (3.6,11,12a).

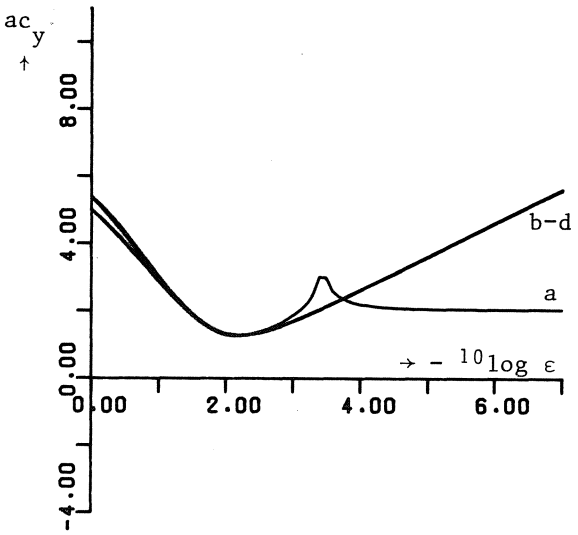

Fig. 3.2 Initial y-error, (3.6,11,12a).

Let us first discuss the results for (3.6,11,12a). Figure 3.1 clearly shows the lack of ε-boudedness of scheme a, i.e., for increasing stiffness its accuracy strongly decreases, whereas the accuracy of b and c remains constant. Also note that, in this case, scheme d is much more accurate than b and c. Figure 3.1 shows that d even takes advantage of increasing stiffness (this phenomenon cannot be explained from the notions of ε-boundedness and ε-accuracy). Figure 3.2 clearly shows the lack of ε-accuracy of scheme a. It should be noted that scheme c, which according to (3.14) is not ε-accurate, yields the same initial y-errors as b and d. This can be explained from the following (heuristic) observation. Consider the linear part of the third component of equation (3.6), i.e., $\dot{y} = -\varepsilon^{-1}\mu(t)y$. Application of the 2-stage schemes c and d to this equation, yields

$$(3.15) \qquad \frac{y_{n+1}}{y_n} = \frac{1-(\lambda_{20}-\gamma)\tau\varepsilon^{-1}\mu(t_n)-(\lambda_{21}-\gamma)\tau\varepsilon^{-1}\mu(t_n+\nu_1\tau)}{(1-\gamma\tau\varepsilon^{-1}\mu(t_n))(1-\gamma\tau\varepsilon^{-1}\mu(t_n+\nu_1\tau))} = O(\varepsilon),\ \varepsilon \to 0.$$

Hence the extra Jacobian evaluation yields extra damping, even if $y_n^{(1)}/y_n = O(1)$.

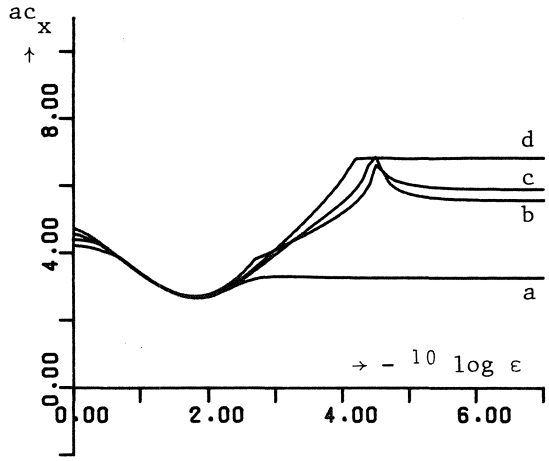

Fig. 3.3 Initial x-error, (3.6,11,12b).

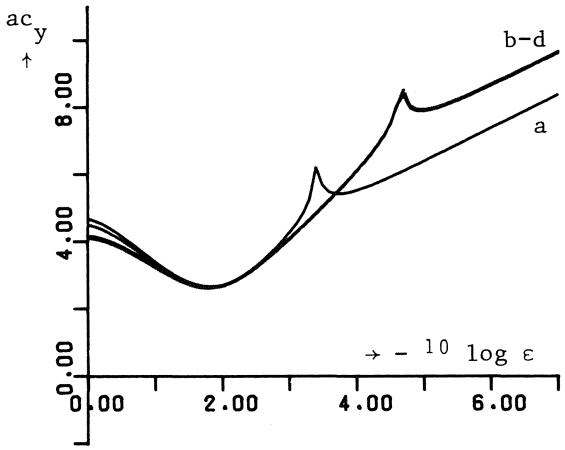

Fig. 3.4 Initial y-error, (3.6,11,12b).

The results for the easier problem (3.6,11,12b) have been plotted in figures 3.3, 3.4. For this problem all x-approximations are $O(1)$ and all y-approximations are $O(\varepsilon)$ (cf. Theorem 3.2, part (v)). Note however that the ε-bounded schemes b-d yield significantly more accuracy than scheme a. Finally it is worthwile to observe that for the larger ε-values, say $\varepsilon \in [10^{-2},1]$, all 4 schemes yield approximately the same errors.

4. MODEL CLASS 2

The second model class we are interested in, and which was also discussed in the previous paper [12], is reviewed in section 4.1. Section 4.2 deals with a specific example which is used in section 4.3 for a numerical illustration.

4.1. The class of D-stability model equations

The following class of linear stiff model problems, class S, was proposed by van Veldhuizen [11] (cf. (2.2)):

$$(4.1.) \qquad \dot{X} = F(t,\varepsilon)X = \varepsilon^{-1}\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} X, \quad X(t) \in \mathbb{C}^2,$$

where

a) $a_{ij} \in \mathbb{C}$ depends smoothly on $t \in [0,T]$ and $\varepsilon \in (0,\varepsilon_0]$.

b) $F(t,\varepsilon) = E(t,\varepsilon)D(t,\varepsilon)E^{-1}(t,\varepsilon)$, where

$$D = \begin{bmatrix} d_1 & 0 \\ 0 & \varepsilon^{-1}d_2 \end{bmatrix}, \quad \mathrm{Re}(d_2(t,\varepsilon)) \leq \tilde{d}_2 < 0 \text{ on } [0,T]\times(0,\varepsilon_0].$$

d_1, d_2, E and E^{-1} depend smoothly on t, ε and the derivatives from order zero up to a sufficiently high order are bounded on $[0,T]\times(0,\varepsilon_0]$.

Van Veldhuizen used class S in his D-stability investigations. Though presented in a somewhat different setting D-stability may be viewed upon as a uniform boundedness property, like ε-boundedness. However, it only applies to linear homogeneous problems $\dot{X} = F(t)X$. For reasons of presentation we therefore do not make use of van Veldhuizen's definition which is slightly different from ours (see [11,12]).

As pointed out in [11], a nice feature of model (4.1) is the possibility to define subclasses of S which describe certain types of couplings between smooth and stiff solution components. Because these couplings may be of decisive importance for the performance of a Rosenbrock type method, we give a short description of these subclasses. Consider a problem from class S. Denote $Y(t) = E^{-1}(t)X(t)$. Then Y satisfies

$$\dot{Y} = [D(t) - C(t)]Y, \quad C(t) = E^{-1}(t)\dot{E}(t). \tag{4.2}$$

In case $C(t)$ is diagonal on $[0,T]$, the problem from S has been uncoupled by the transformation $X = EY$, i.e., there exists no coupling between smooth and transient components. Otherwise we employ

DEFINITION 4.1. The coupling from the smooth to the transient component, at $t = t^*$, is weak if $C_{21}(t^*) = O(\varepsilon)$. The coupling from the transient to the smooth component, at $t = t^*$, is weak if $C_{12}(t^*) = O(\varepsilon)$. If a coupling is not weak, we call it strong. W_{st} (W_{ts}) denotes the subclass of S for which on the whole time interval $C_{21}(t) = O(\varepsilon)$ ($C_{12}(t) = O(\varepsilon)$). □

Due to assumptions (4.1a,b) the matrix C(t) is at least O(1) as $\varepsilon \to 0$. Hence problem (4.2) is of type (3.1). By means of Theorem 3.1, and the bounded transformation X = EY, it thus follows that all solutions of (4.1) are bounded in $\varepsilon \in (0,\varepsilon_0]$.

THEOREM 4.1. *Consider an arbitrary 2-stage Rosenbrock method* (1.2) *which evaluates* J(t,X) *once per integration step. This method is*
(i) *ε-bounded on W_{ts}.*
(ii) *ε-bounded on W_{st}, iff* $R^{(1)}(\infty) = 0$.
(iii) *not ε-bounded on S.*

PROOF. This theorem is a special case of Theorem 3.1 in [11]. □

THEOREM 4.2. *An* m*-stage Rosenbrock method* (1.2) *is ε-bounded on S iff at each stage* J(t,X) *is evaluated at the special point* $(t,X) = (t^{(j)},X^{(j)})$.

PROOF. The necessity follows from Theorem 4.1 , part (iii). Recall that boundedness of the m-stage result implies, by definition, boundedness of the preceding m-1 results. The sufficiency has been proved in [12], Theorem 3.1. □

These 2 theorems show that if we have a strong coupling from stiff to smooth, and vice versa, ε-boundedness cannot be guaranteed if we restrict ourselves to one J(t,x)-evaluation per integration step. Unfortunately, schemes which reevaluate the Jacobian per stage are usually not recommended because of their considerable computational overhead.

So far we did not yet attempt to prove part (i) and (ii) of Theorem 4.1 for methods (1.2) using more than 2 stages. We do conjecture however that these methods are also ε-bounded on W_{ts}, and ε-bounded on W_{st}, iff $R^{(j)}(\infty) = 0$, $j < m$. For example, the class consisting of all problems

$$\dot{X} = \begin{bmatrix} a_{11} & \varepsilon^{-1}a_{12} \\ a_{21} & \varepsilon^{-1}a_{22} \end{bmatrix} X \tag{4.3}$$

satisfying properties (4.1a,b), is a subclass, say S_2, of W_{st} [12]. Because (4.3) may also be viewed upon as a prototype of the first variational form

of model (3.1), part (iv) of Theorem (3.2) applies. It thus follows that an m-stage Rosenbrock method (1.2), using one J(t,X)-evaluation per step, is ε-bounded on S_2, iff $R^{(j)}(\infty) = 0$ for $j < m$.

Because $S_2 \subset W_{st}$, class S_2 does not describe strong couplings from smooth to transient. This fact may be considered as a shortcoming of equation (4.3), and thus also of (3.1), when used as a model.

4.2. A test example exhibiting only strong couplings

Consider the problem (see also [6,8,12])

$$\dot{X} = E(t) \begin{bmatrix} d_1(t) & 0 \\ 0 & \varepsilon^{-1} d_2(t) \end{bmatrix} E^{-1}(t)X, \quad E(t) = \begin{bmatrix} \cos\theta t & -\sin\theta t \\ \sin\theta t & \cos\theta t \end{bmatrix}, \tag{4.4}$$

θ being constant. Then $Y(t) = E^{-1}(t)X(t)$ satisfies (cf. (4.2))

$$\dot{Y} = \begin{bmatrix} d_1(t) & \theta \\ -\theta & \varepsilon^{-1} d_2(t) \end{bmatrix} Y. \tag{4.5}$$

Hence $C_{12}(t) = -\theta$, $C_{21}(t) = \theta$. Consequently, we have to deal with a strong coupling from stiff to smooth, and vice versa. It is not difficult to verify that for this specific example part (iii) of Theorem 4.1 applies. Note that equation (4.5) belongs to S_2. Let $d_1 = d_2 = -1$. Then

$$Y(t) = \begin{bmatrix} 1+\varepsilon\lambda^+ & 1+\varepsilon\lambda^- \\ -\varepsilon\theta & -\varepsilon\theta \end{bmatrix} \begin{bmatrix} C^+ e^{\lambda^+ t} \\ C^- e^{\lambda^- t} \end{bmatrix}, \tag{4.6}$$

where $C^{\pm}$ are arbitrary constants and $\lambda^{\pm} = \frac{1}{2}(-1-\varepsilon^{-1} \pm \sqrt{(1-\varepsilon^{-1})^2 - 4\theta^2})$. Note that $\lambda^- \sim -\varepsilon^{-1}$ and $\lambda^+ \to -1$ as $\varepsilon \to 0$. Next we set $C^- = 0$, $C^+ = 1$. Then

$$X(t) = E(t) \begin{bmatrix} (1+\varepsilon\lambda^+)e^{\lambda^+ t} \\ -\varepsilon\theta e^{\lambda^+ t} \end{bmatrix} = \begin{bmatrix} e^{-t}\cos\theta t \\ e^{-t}\sin\theta t \end{bmatrix} + O(\varepsilon), \quad \varepsilon \to 0. \tag{4.7}$$

We see that solution (4.7) is smooth and, to a great extent, independent of the stiffness parameter ε. The same remark applies to the first component of the corresponding solution of (4.5). Its second somponent is $O(\varepsilon)$. In what follows we shall refer to the X-example and Y-example.

4.3. Numerical illustration

We integrated the X-example and Y-example for $\theta = 1$ and for a set of ε-values from $[10^{-8}, 10^{-1}]$ with all 4 two-stage formulas (3.14) over the t-interval $[0, 2\pi]$, using a constant stepsize $\tau = \pi/25$. Note that for the Y-example the formulas (3.14) are identical.

In figure 4.1 we plotted the value ac = $-^{10}\log$(max. abs. error at t = 2π) against ε. The a-curve and b-curve clearly show the lack of ε-boundedness of methods a and b (instability for small ε). Methods c and d are ε-bounded on S (see Theorem 4.2). They produce approximations which are nearly independent of ε. Recall that, for small ε, the exact solutions share this property. Finally, this example nicely shows that a simple transformation of the differential equation may lead to a qualitatively different behaviour of a Rosenbrock method.

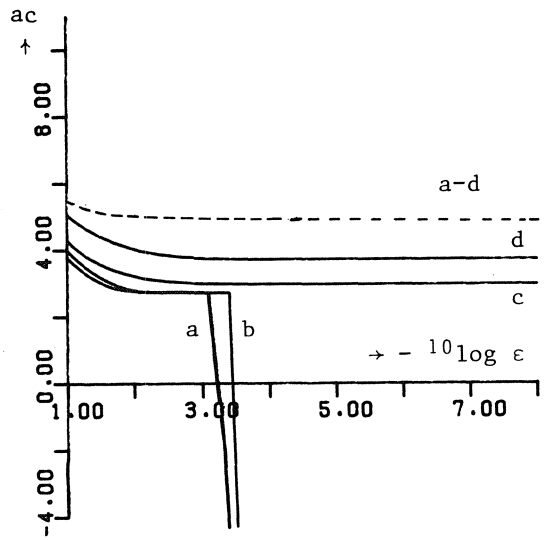

Fig. 4.1 ——— X-example;
---- Y-example.

5. THE AUTONOMOUS NOTATION

Many authors prefer the autonomous notation. It facilitates the analysis of the consistency conditions, while every non-autonomous equation (1.1) can be converted to the autonomous form by introducing t as a new dependent variable. For example, the Rosenbrock code ROW4A requires the autonomous

form [1]. When we rewrite problem (1.1) to the autonomous form the derivative F_t enters into the computation. It is easily seen that the Rosenbrock approximation (1.2) then can be defined by the (non-autonomous) scheme

$$
\begin{aligned}
&X_n^{(0)} = X_n, \\
&K_n^{(j)} = [I-\gamma_j\tau J_n^{(j)}]^{-1}[F(t_n^{(j)}, X_n^{(j)}) + \gamma_j\tau G_n^{(j)}], \quad j = 0(1)m-1, \\
&X_n^{(j)} = X_n + \tau \sum_{\ell=0}^{j-1} \lambda_{j,\ell} K_n^{(\ell)}, \quad j = 1(1)m, \\
&X_{n+1} = X_n^{(m)},
\end{aligned}
\tag{5.1}
$$

where $G_n^{(j)} = G(\hat{t}_n^{(j)}, \hat{X}_n^{(j)})$, $G(t,X) = \partial F(t,X)/\partial t$. Furthermore, $t_n^{(j)}$ is now defined by $t_n^{(j)} = t_n + \tau(\lambda_{j,0} + \ldots + \lambda_{j,j-1})$. All other quantities are defined as in scheme (1.2). It is convenient to use notation (5.1) (cf. [5,9]).

Because we deal with non-autonomous models, the following interesting question arises. When we apply (5.1) to the model classes (3.1) and (4.1), do we then preserve the boundedness results summarized in the 2 preceding sections? For the most interesting results the answer to this question is, peculiarly, negative. It is even negative for schemes using more than 1 Jacobian evaluation per step. This matter will be discussed in section 5.1. By way of illustration, we also repeat the experiments presented before. Section 5.2 reports an experiment with the automatic code ROW4A.

5.1. Boundedness results for method 5.1

THEOREM 5.1. *No Rosenbrock method* (5.1) *is* ε*-accurate on class* (3.1).

PROOF. By counterexample. Consider the simplified problem $\dot{y} = -\varepsilon^{-1}\mu(t)y$. Application of any method (5.1), at a point (t,y), delivers

$$
y^{(1)} = \frac{1-(\lambda_{10}-\gamma_0)\tau\varepsilon^{-1}\mu(t) - \lambda_{10}\gamma_0\tau^2\varepsilon^{-1}\dot{\mu}(t)}{1+\gamma_0\tau\varepsilon^{-1}\mu(t)}\, y. \tag{5.2}
$$

We see that if $(\lambda_{10}-\gamma_0)\mu(t) + \lambda_{10}\gamma_0\tau\,\dot{\mu}(t) \neq 0$, then $y^{(1)} = O(1)$ as $\varepsilon \to 0$. By definition, ε-accuracy of an m-stage method implies $y^{(1)} = O(\varepsilon)$. □

If we apply relation (5.2) repeatedly, we may easily encounter instability. For example, substituting $\lambda_{10} = \gamma_0$ (L-stability) and $\mu(t) = \exp(-t/\tau\gamma_0)$ yields $y^{(1)} = y$. On the other hand, when using the non-autonomous notation the substitution $\lambda_{10} = \gamma_0$ delivers $|y^{(1)}/y| = |(1+\gamma_0\tau\,\varepsilon^{-1}\mu(t))^{-1}| < 1$ for all $\tau > 0$ and $\varepsilon \in (0,\varepsilon_0]$. In other words, the stability of the 1-stage scheme may be lost by conversion to the autonomous form. Without doubt this conclusion also applies to m-stage schemes, $m > 1$. As we do not discuss stability properties we do not pursue this subject further.

THEOREM 5.2. (i) *No method* (5.1) *is* ε*-bounded on class* S_2. *Consequently, no method* (5.1) *is* ε*-bounded on class* S *and class* (3.1).
(ii) *Any method* (5.1) *is* ε*-bounded on the* 2 *classes of problems* (3.1) *for which, respectively,* $A = 0$ *and* A,μ *are constant.*
(iii) *Consider class* (3.1). *Let the point* $X = (x,y)$ *occurring in Definition* (2.1) *be such that* $x = O(1)$ *and* $y = O(\varepsilon)$. *Then any Rosenbrock method* (5.1) *is* ε*-accurate on the whole class* (3.1).

PROOF. The proofs of (ii) - (iii) go along the same lines as the proofs of the corresponding parts of Theorem 3.2 (see [12], section 4). The proof of part (i) goes by counter-example. It suffices to take $m = 1$. Consider the problem (cf. (4.3))

$$\begin{aligned} \dot{x} &= \varepsilon^{-1}a_{12}(t)y, \\ \dot{y} &= -\varepsilon^{-1}y. \end{aligned} \tag{5.3}$$

The 1-stage scheme, applied at a point (t,x,y), yields the increment vector

$$K^{(0)} = \begin{bmatrix} \varepsilon^{-1}a_{12}(t)(1+\gamma_0\tau\,\varepsilon^{-1})^{-1}y + \gamma_0\tau\,\varepsilon^{-1}\dot{a}_{12}(t)y \\ -\varepsilon^{-1}(1+\gamma_0\tau\,\varepsilon^{-1})y \end{bmatrix}. \tag{5.4}$$

By an appropriate choice of $a_{12}(t)$, the first component becomes unbounded in $\varepsilon \in (0,\varepsilon_0]$. This simple observation proves part (i). □

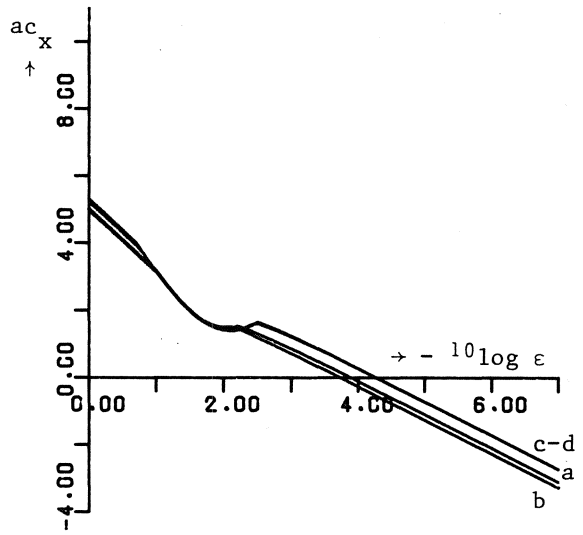

Fig. 5.1 Initial x-error, (3.6,11,12a). Autonomous notation

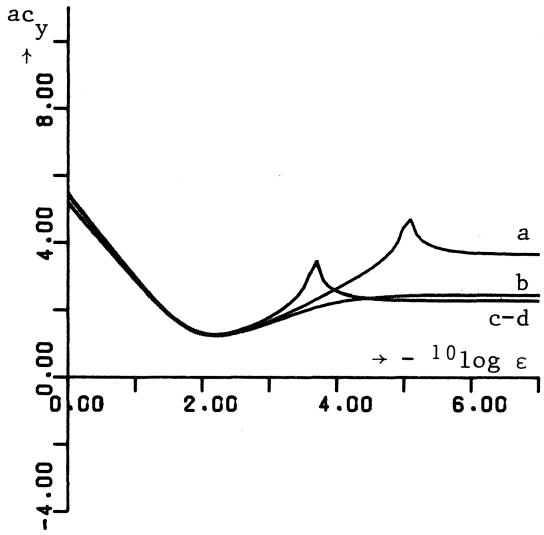

Fig. 5.2 Initial y-error, (3.6,11,12a). Autonomous notation

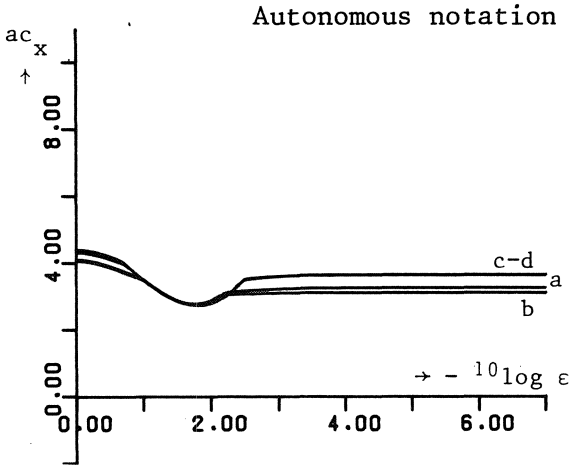

Fig. 5.3 Initial x-error, (3.6,11,12b). Autonomous notation

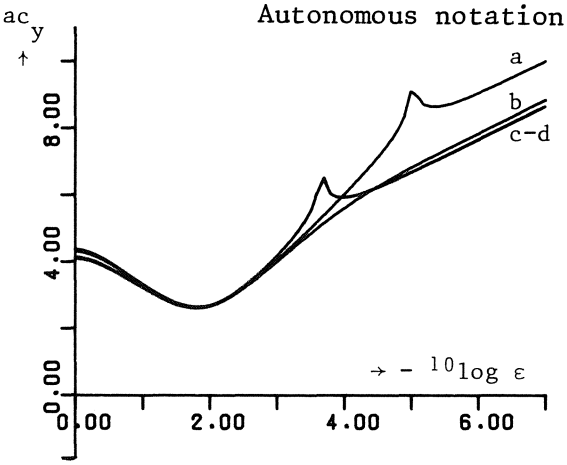

Fig. 5.4 Initial y-error, (3.6,11,12b). Autonomous notation

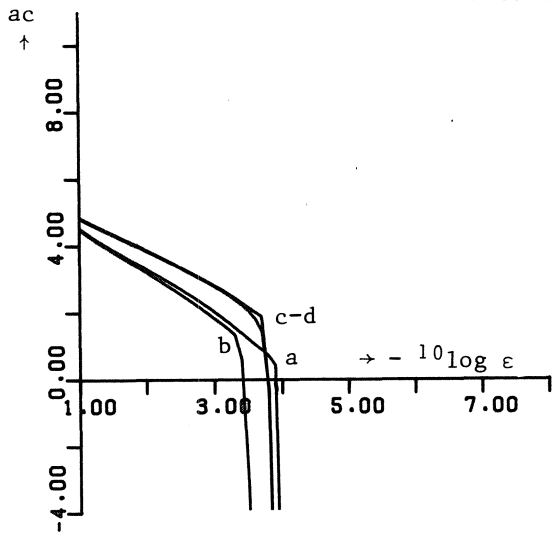

Fig. 5.5 X-example. Autonomous notation

By way of illustration we repeated the afore-mentioned experiments with the 4 two-stage schemes (3.14), but now using the autonomous form. Figures 5.1-5.4 and 5.5 correspond with figures 3.1-3.4 and 3.5, respectively. Note that all 4 schemes now behave more or less equal.

5.2. An experiment with ROW4A

ROW4A is an automatic Rosenbrock code based on the algorithm GRK4A published in [4]. Gottwald and Wanner [1] provided it with a so-called back-step strategy to obtain a more reliable stepsize and local error control. The underlying integration method is A-stable and of order 4. Its increment vectors are of the (more general) form (cf. (1.2))

$$(5.5) \qquad K_n^{(j)} = [I - \gamma\tau J_n]^{-1}\{F(X_n^{(j)}) + \sum_{\ell=0}^{j-1} \beta_{j,\ell}\, K_n^{(\ell)}\},$$

and assume the autonomous notation ((5.5) can also be rewritten like formula (5.1), see [5]). The method is not ε-bounded on class (3.1) and class S. It uses 3 F(X)-evaluations and 1 J(X)-evaluation per step. We implemented ROW4A on a CDC Cyber 750 in single precision (14 decimals). Our version computes J_n from the analytic expression.

Our aim of reporting an experiment with an automatic code, like ROW4A, is to illustrate how the lack of ε-boundedness shows up in practice. When this property is missing, one may encounter unusually large local errors, even when the solution to be integrated is smooth. A reliable code should detect these errors and should, at the cost of the number of integration steps of course, deliver a result of the desired accuracy (see also [11], section 5). In view of this, ROW4A seems to suit our purpose as it has been equipped with the back-step strategy.

The experiment consists of the automatic integration of the X-example and Y-example of section (4.3), over the interval $[0,2\pi]$, for a set of ε-values between 10^{-7} and 10^{-1}. The tolerance parameter TOL of ROW4A and the initial stepsize were in all integrations equal to 10^{-3} and 10^{-2}, respectively. Figure 5.6 shows results of the experiment.

The plots clearly show the lack of ε-boundedness of ROW4A when applied to the X-example. Though the exact solution is smooth, and nearly independent of ε, the numbers IPAS and IREP strongly increase as ε decreases. As observed above, such a behaviour was to be expected. However, more dramatic is that the code looses its accuracy. The local error control clearly fails on this example. This experiment confirms that it can be very dangerous to rely on local error control mechanisms.

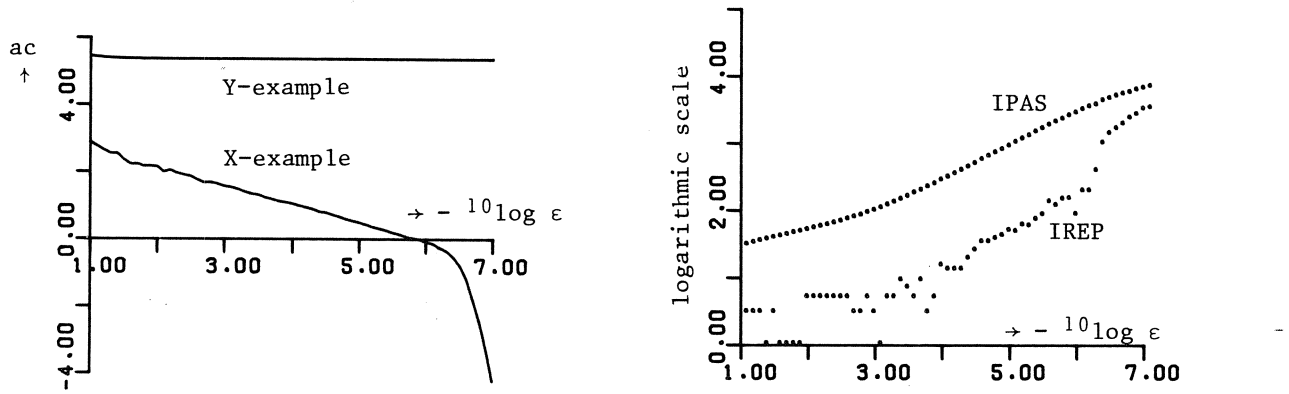

Fig. 5.6 Results for ROW4A. In the right figure we plotted IPAS = the number of accepted steps and IREP = the number of repeated steps needed by ROW4A on the X-example. For the Y-example these numbers are 16 and 0, respectively, and do not change with ε.

6. SOME FINAL COMMENTS

The question arises how to employ our experiences in order to improve the Rosenbrock methods when applied to real life problems. Let us first consider methods based on the *non-autonomous* notation (1.1). For this type of Rosenbrocks methods our results strongly suggest to take care of ε-boundedness and ε-accuracy when dealing with problems where the stiffness originates from t-dependent parts in the equation. However, if one wishes to construct such a method, one has to face an additional difficulty, i.e., the solution of extra order conditions due to the presence of derivatives to t. To solve these extra conditions, for a given order, it may well be necessary to add extra stages. From this point of view the *autonomous* notation should be preferred. Unfortunately, for the type of problems mentioned above *the conversion to the autonomous form* may lead to a significant loss in accuracy, as shown in our experiments. This circumstance makes it difficult to decide which approach should be preferred. In the author's opinion, an accountable decision can only be made if one has a typical problem class at hand. In this connection we should also remark that Kaps and Rentrop [4] and Gottwald and Wanner [1] report promising results with their 'autonomous' codes GRK4A and ROW4A. Gottwald and Wanner [2] even show that on a set of 4

real life problems from chemical kinetics and physiology, their code ROW4A is more efficient and more reliable than a popular backward differentiation one.

ACKNOWLEDGEMENT.

The author would like to express his thanks to Mrs. M.J. Louter-Nool for her assistance in preparing the plots.

REFERENCES

[1] GOTTWALD, B.A. & G. WANNER, *A reliable implementation of one-step methods for differential equations*, Computing, to appear.

[2] ——, *Stiff systems of ordinary differential equations in biology and chemistry: Validation of numerical methods for their solution* in: Continuous Simulation of Physical Systems, T.D. Bui (ed.), to appear.

[3] GRIEPENTROG, E., *Numerische Integration steifer Differential-gleichungssysteme mit Einschrittverfahren*, Beiträge zur Numerischen Mathematik 8, 59-74 (1980).

[4] KAPS, P. & P. RENTROP, *Generalized Runge-Kutta methods of order 4 with stepsize control for stiff ordinary differential equations*, Numer. Math. 33, 55-68 (1979).

[5] KAPS, P. & G. WANNER, *A study of Rosenbrock type methods of high order*, Report Universität Innsbruck (1979).

[6] KREISS, H.O., *Difference methods for stiff ordinary differential equations*, SIAM J. Numer. Anal. 15, 21-58 (1978).

[7] ROSENBROCK, H.H., *Some general implicit processes for the numerical solution of differential equations*, Computer J. 5, 329-330 (1963).

[8] SAND, J., *A note on a differential system constructed by H.O. Kreiss*, Report TRITA-NA-8004, The Royal Institute of Technology, Stockholm, Sweden (1980).

[9] SHAMPINE, L.F., *Implementation of Rosenbrock Methods*, Report SAND80-2367J, Sandia National Laboratories, Albuquerque, New Mexico (1980).

[10] STETTER, H,J., *Towards a theory for discretizations of stiff differential systems*, Lecture Notes in Mathematics 506, Springer Verlag, Berlin, 190-201 (1976).

[11] VELDHUIZEN, M. VAN, D-*stability*, SIAM J. Numer. Anal., to appear.

[12] VERWER, J.G., *An analysis of Rosenbrock methods for non-linear stiff initial value problems*, SIAM J. Numer. Anal., to appear.