Centrum voor Wiskunde en Informatica

REPORT*RAPPORT*

MAS

Modelling, Analysis and Simulation

Modelling, Analysis and Simulation

On global error estimation and control for initial value problems

J. Lang, J.G. Verwer

REPORT MAS-E0531 DECEMBER 2005

CWI is the National Research Institute for Mathematics and Computer Science. It is sponsored by the Netherlands Organization for Scientific Research (NWO).
CWI is a founding member of ERCIM, the European Research Consortium for Informatics and Mathematics.

CWI's research has a theme-oriented structure and is grouped into four clusters. Listed below are the names of the clusters and in parentheses their acronyms.

Probability, Networks and Algorithms (PNA)

Software Engineering (SEN)

Modelling, Analysis and Simulation (MAS)

Information Systems (INS)

P.O. Box 94079, 1090 GB Amsterdam (NL)
Kruislaan 413, 1098 SJ Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

ISSN 1386-3703

On global error estimation and control for initial value problems

ABSTRACT
This paper addresses global error estimation and control for initial value problems for ordinary differential equations. The focus lies on a comparison between a novel approach based on the adjoint method combined with a small sample statistical initialization and the classical approach based on the first variational equation. Control is achieved through tolerance proportionality. Both approaches are found to work well and to enable estimation and control in a reliable manner. However, the novel approach is not found to be competitive with the classical approach, mainly because of its huge storage demand for large problems.

2000 Mathematics Subject Classification: Primary: 65L05, 65L06, 65L20, 65L70, 65M20
1998 ACM Computing Classification System: G.1.7, G.1.8
Keywords and Phrases: Numerical integration for ODEs and PDEs; global error estimation; global error control; defects and local errors; tolerance proportionality and local errors; adjoint method; small sample statitistical initialization
Note: Work carried out within theme MAS1

On Global Error Estimation and Control for Initial Value Problems

J. Lang
Darmstadt University of Technology
Schlossgartenstrasze 7, 64289 Darmstadt, Germany
lang@mathematik.tu-darmstadt.de

J.G. Verwer
Center for Mathematics and Computer Science
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands
Jan.Verwer@cwi.nl

December 14, 2005

Abstract

This paper addresses global error estimation and control for initial value problems for ordinary differential equations. The focus lies on a comparison between a novel approach based on the adjoint method combined with a small sample statistical initialization and the classical approach based on the first variational equation. Control is achieved through tolerance proportionality. Both approaches are found to work well and to enable estimation and control in a reliable manner. However, the novel approach is not found to be competitive with the classical approach, mainly because of its huge storage demand for large problems.

2000 Mathematics Subject Classification: Primary: 65L05, 65L06, 65L20, 65L70, 65M20.
1998 ACM Computing Classification System: G.1.7, G.1.8.
Keywords and Phrases: Numerical integration for ODEs and PDEs, global error estimation, global error control, defects and local errors, tolerance proportionality, adjoint method, small sample statistical initialization.
Note: Work carried out within theme MAS1.

1 Introduction

Suppose one is given in $\mathbb{R}^m$ the initial value problem for the system of ODEs

$$w' = F(t,w)\,, \quad w(0) = w_0\,, \qquad 0 < t \le T\,, \tag{1.1}$$

and a sequence of approximations w_n to its exact solution values $w(t_n)$ computed by a numerical integration method at a certain time grid

$$0 = t_0 < t_1 < \cdots < t_n < \cdots < t_{N-1} < t_N = T\,. \tag{1.2}$$

Hereby the major research concerns are efficiency: how to get the w_n at minimal CPU costs (if $m \gg 1$), and reliability: how large are the global errors

$$\varepsilon_n = w(t_n) - w_n\,, \qquad n = 0\,, \ldots\,, N\,. \tag{1.3}$$

In the past, numerical ODE research has focused on the efficiency question. The reliability question has received much less attention in spite of the by now twenty years old survey paper 'Thirteen ways to estimate global error' [20]. Existing popular codes focus on efficiency by adaptively optimizing time grids (1.2) in accordance with local error control. Such a control makes sense if solutions exhibit sharp changes at local intervals much smaller than the total interval $[0, T]$ and are smooth elsewhere. However, local errors (errors made within a single integration step) may substantially differ from the global ones (1.3). This largely depends on the conditioning (stability) of the system (1.1) at hand (sensitivity to growth in time of perturbations of w_0 and $F(t, w)$). If a system is well-conditioned, a well designed local error control [17, 18] will work out reliably. But if the conditioning is bad, even the best designed local error control should not be trusted.

For global error control it is necessary to take into account the conditioning of system (1.1), similar to the matrix condition number in numerical linear algebra. Herewith it is desirable to avoid strict a priori error bounds as these can be overly pessimistic, e.g. when fortunate cancellation effects occur. Taking into account the conditioning of system (1.1) is best done during actual computation. This requires at least two full integrations over $[0, T]$, both in the classical (forward-forward) and the adjoint (forward-backward) approach. For large-scale ODE problems, such as spatially discretized multi-dimensional PDEs, a clear disadvantage of the latter is that many or all approximations w_n computed on the whole of (1.2) must be stored. Yet, in recent years the adjoint approach has gained popularity to obtain goal oriented a posteriori error estimation for an adaptive error control for PDEs [1, 5, 7] and ODEs [2, 6, 13].

Our interest in this paper lies in estimation and control of the global errors (1.3). Specifically, regarding estimation we will compare a novel approach based on the adjoint method combined with a small sample statistical initialization proposed by Cao & Petzold [2] to the classical approach based on the first variational equation. For both, global control is achieved by exploiting the property of tolerance proportionality derived from local control [17]. As a typical integration method we will use the existing Runge-Kutta-Rosenbrock method ROS3P [16]. Both approaches are found to work well and to enable estimation and control in a reliable manner. However, the novel approach is not found to be competitive with the classical approach, mainly because of its huge storage demand for large problems.

2 The classical (forward-forward) approach

2.1 The perturbed system

In the spirit of backward error analysis and following an approach proposed in [23] (see also [20, 21]), let us suppose that there exists a nearby solution $v(t) \approx w(t)$ being the exact solution of a perturbed system

$$v' = F(t, v) + r(t)\,, \quad v(0) = v_0\,, \qquad 0 < t \le T\,, \tag{2.1}$$

with perturbation $v_0 - w_0$ at $t = 0$ and perturbation $r(t)$ for $0 < t \le T$. Assuming F to be continuously differentiable, so that the mean-value theorem for vector functions is

applicable, the error function

$$e(t) = v(t) - w(t)\,, \qquad 0 \le t \le T\,, \tag{2.2}$$

then satisfies

$$e' = A(t)e + r(t)\,, \quad e(0) = v_0 - w_0\,, \qquad 0 < t \le T\,, \tag{2.3}$$

where

$$A(t) = \int_0^1 F'\big(t, v(t) + (s-1)e(t)\big)ds = \int_0^1 F'\big(t, w(t) + se(t)\big)ds\,. \tag{2.4}$$

Here F' denotes the Jacobian matrix with respect to the dependent variable. As we speak of perturbations we tacitly assume that $e(0)$ is significantly smaller than w_0 and likewise that $r(t)$ is significantly smaller than $F\big(t, v(t)\big)$. The error function $e(t)$ can be expressed as

$$e(t) = \Phi^{-1}(t)\Phi(0)\,e(0) + \Phi^{-1}(t)\int_0^t \Phi(s)r(s)\,ds\,, \tag{2.5}$$

where the $m \times m$ fundamental matrix solution $\Phi(t)$ solves $\Phi'(t) = -\Phi(t)A(t)$. Note that $\Phi(t) = \Phi(0)\exp(-At)$ for constant A where $\Phi(0)$ is an arbitrary nonsingular matrix. Apparently, the product $\Phi^{-1}(t)\Phi(s), 0 \le s \le t$, governs growth or decay of $e(t)$ in time and thus determines the conditioning of (1.1).

Remark 2.1 For practical purposes the fundamental matrix solution concept is clearly of little use. An a priori bound which sometimes can be used in practice is based on the logarithmic matrix norm. Let $\mu\big(A(t)\big)$ denote the logarithmic matrix norm of $A(t)$ and suppose $\mu\big(A(t)\big) \le \omega$ with ω a constant valid for all $t \in [0, T]$. Then, with $\|\cdot\|$ denoting the associated vector norm in $\mathbb{R}^m$, there holds [3, 12]

$$\|e(t)\| \le e^{\omega t}\|e(0)\| + \int_0^t e^{\omega(t-s)}\|r(s)\|ds\,, \qquad 0 \le t \le T\,. \tag{2.6}$$

This a priori bound is useful for strictly negative ω. Let for simplicity $e(0) = 0$ and suppose that $\|r(t)\|$ is uniformly bounded by a small number δ_r. Then

$$\|e(t)\| \le \frac{1}{\omega}\Big(e^{\omega t} - 1\Big)\delta_r\,, \tag{2.7}$$

uniformly bounding $e(t)$ in norm by $\delta_r/|\omega|$ for infinite time if $\omega < 0$. Due to the negative logarithmic norm this is an instance of a well-conditioned ODE system (1.1). On the other hand, with a positive logarithmic norm the bound predicts exponential growth in time. Then we speak of an ill-conditioned ODE system, provided the bound is realistic or even sharp. However, any result of this sort depends on the chosen vector norm since $\mu\big(A(t)\big)$ depends on the norm. More precisely, it can happen that $\mu\big(A(t)\big) \gg 0$, even if in a global sense the problem is well conditioned, see e.g. [3] and Example I.2.5 in [12]. Also, for hard problems from practice, finding estimates of ω may be very cumbersome. $\diamond$

For properly using the perturbed system in the classical (forward-forward) approach for p-th order consistent one-step integration methods, it is helpful to associate equation (2.3)

with the well-known first variational equation for the global error ε_n of such a one-step method. For this purpose we adopt for the one-step method the general Henrici notation [9]

$$w_{n+1} = w_n + \tau_n \Psi(t_n, w_n; \tau_n)\,, \qquad \tau_n = t_{n+1} - t_n\,, \tag{2.8}$$

and introduce its local error

$$\delta_n = w(t_{n+1}) - w(t_n) - \tau_n \Psi(t_n, w(t_n); \tau_n)\,. \tag{2.9}$$

Assuming now a constant step size τ (for simplicity only), p-th order consistency and a sufficiently smooth solution, the local error defined at $\big(t, w(t)\big)$

$$\delta(t) = w(t+\tau) - w(t) - \tau \Psi(t, w(t); \tau)\,, \tag{2.10}$$

possesses an expansion of the form $\delta(t) = \tau\rho(t) + \mathcal{O}(\tau^{p+2})$. Hence in this notation the principal error function $\rho(t) = \mathcal{O}(\tau^p)$. Next, let $G(t) = F'\big(t, w(t)\big)$. With this local expansion at hand the global error ε_n is then known to possess an expansion $\varepsilon_n = \eta(t_n) + \mathcal{O}(\tau^{p+1})$ where $\eta(t)$ is the solution of the first variational equation

$$\eta' = G(t)\eta + \rho(t)\,, \quad \eta(0) = 0\,, \qquad 0 < t \le T\,. \tag{2.11}$$

We refer to [9], Section 3.3 and [8], Section II.8 for further details.[1)] One can now connect equation (2.3) to (2.11). To that end put $e(0) = 0$ (this bears no restriction) and approximate $A(t) = G(t) + \mathcal{O}\big(e(t)\big)$ by $G(t)$, so that (2.3) is replaced by

$$e' = G(t)e + r(t)\,, \quad e(0) = 0\,, \qquad 0 < t \le T\,. \tag{2.12}$$

Apparently, by implementing a proper choice of the defect $r(t)$, solving (2.3) or likewise (2.12) will in leading order amount to solving the first variational equation (2.11) for the true global error. We will illustrate this next.

2.2 Interpolation and defect computation

We define the nearby solution $v(t)$ by piecewise cubic Hermite interpolation of the given approximation sequence. Hence at every subinterval $[t_n, t_{n+1}]$, $n = 0, 1, \ldots, N-1$, we form

$$P(t) = w_n + (t - t_n)\,\mathcal{A}_n + (t - t_n)^2\,\mathcal{B}_n + (t - t_n)^3\,\mathcal{C}_n\,, \qquad t_n \le t \le t_{n+1}\,, \tag{2.13}$$

and choose the coefficients such that $P(t_n) = w_n, P(t_{n+1}) = w_{n+1}$ and $P'(t_n) = F_n$, $P'(t_{n+1}) = F_{n+1}$ where $F_n = F(t_n, w_n)$ and $F_{n+1} = F(t_{n+1}, w_{n+1})$. This gives

$$\begin{aligned} \mathcal{A}_n &= F_n\,, \\ \mathcal{B}_n &= \big(3w_{n+1} - 3w_n - \tau F_{n+1} - 2\tau F_n\big)/\tau^2\,, \\ \mathcal{C}_n &= \big(2w_n - 2w_{n+1} + \tau F_n + \tau F_{n+1}\big)/\tau^3\,. \end{aligned} \tag{2.14}$$

Would one interpolate (smooth) exact solution values $w(t), w'(t)$ at $t = t_n$ and $t = t_{n+1}$, then

$$\mathcal{A} = w^{(1)}\,, \quad \mathcal{B} = \frac{1}{2}w^{(2)} - \frac{1}{24}\tau^2 w^{(4)} + \mathcal{O}(\tau^5)\,, \quad \mathcal{C} = \frac{1}{6}w^{(3)} + \frac{1}{12}\tau w^{(4)} + \mathcal{O}(\tau^5)\,, \tag{2.15}$$

[1)] In spite of the fact that the asymptotic theory behind (2.11) is classical in the sense that F is assumed Lipschitz without taking into account stiffness ($\tau\|F'(w)\| \gg 1$), we will use it for stiff problems in Section 4.

where the expressions are evaluated at $t = t_n$. Writing $t = t_n + s\tau$, $0 \le s \le 1$, yields

$$\begin{aligned} P(t) &= w(t) + \tfrac{1}{24}\big(2s^3 - s^2 - s^4\big)\tau^4 w^{(4)}(t_n) + \mathcal{O}(\tau^5)\,, \\ P'(t) &= w'(t) + \tfrac{1}{24}\big(6s^2 - 2s - 4s^3\big)\tau^3 w^{(4)}(t_n) + \mathcal{O}(\tau^4)\,, \end{aligned} \tag{2.16}$$

so that at every subinterval $[t_n, t_{n+1}]$ the defect function $d(t) = P'(t) - F\big(t, P(t)\big)$ satisfies

$$d(t) = P'(t) - w'(t) + F\big(t, w(t)\big) - F\big(t, P(t)\big) = \begin{cases} \mathcal{O}(\tau^3)\,, & s \neq \frac{1}{2}\,, \\ \mathcal{O}(\tau^4)\,, & s = \frac{1}{2}\,. \end{cases} \tag{2.17}$$

Here we have assumed that F is Lipschitz and have used that $6s^2 - 2s - 4s^3 = 0$ for $s = \frac{1}{2}$.

In actual application the interpolation is based on numerical approximations of order p. By assuming exact local solution values at $t = t_n$ and corresponding local Taylor expansions of w_{n+1}, F_{n+1} it then follows in the same way as above that $d(t) = \mathcal{O}(\tau^q)$ where $q = \min(p, 3)$, and with the special value $s = \frac{1}{2}$ we have $q = \min(p, 4)$. Hence the cubic Hermite polynomial can be used up to consistency order $p = 3$, and when using only $s = \frac{1}{2}$ even up to order $p = 4$. In the remainder we will employ the defect $d(t)$ halfway the step intervals, that is,

$$d(t_{n+1/2}) = \frac{3w_{n+1} - 3w_n}{2\tau} - \frac{F_n + F_{n+1}}{4} - F\Big(t_{n+1/2}, \frac{w_n + w_{n+1}}{2} + \frac{\tau}{8}\left(F_n - F_{n+1}\right)\Big)\,. \tag{2.18}$$

Now, let $w_n = w(t_n)$ and consider local expansions

$$w_{n+1} = w + \tau w' + \sum_{k=2}^{p} \frac{\tau^k}{k!} w^{(k)} + \frac{\tau^{p+1}}{(p+1)!} C_{p+1} + \mathcal{O}(\tau^{p+2})\,, \tag{2.19}$$

with the righthand side evaluated at $w = w(t_n)$ and an empty sum for $p = 1$. Hence the ρ-term in the local error expansion at $t = t_n$ then reads

$$\rho_n = \frac{\tau^p}{(p+1)!}\Big(w^{(p+1)}(t_n) - C_{p+1}\Big)\,, \tag{2.20}$$

and inserting the expansion for w_{n+1} into $d(t_{n+1/2})$ will reveal

$$d(t_{n+1/2}) = -\frac{3}{2}\,\rho_n + \mathcal{O}(\tau^{p+1})\,, \qquad 1 \le p \le 3\,. \tag{2.21}$$

The cubic Hermite defect halfway the step interval thus can be used to retrieve in leading order the local error of any one-step method of order $1 \le p \le 3$.

Finally we connect (2.12) and (2.11) by putting $r(t_{n+1/2}) = -\frac{2}{3}\,d(t_{n+1/2})$ in the stepwise frozen version of (2.12), i.e.,

$$e' = F'(t_n, w_n)e + r(t_{n+1/2})\,, \quad t_n < t \le t_{n+1}\,, \qquad n = 0, \ldots, N-1\,, \tag{2.22}$$

which will be integrated for the global error estimation. In this manner we actually work in leading order with the stepwise frozen version of the first variational equation (2.11) for the true global error (both $G(t)$ and $\rho(t)$ are frozen at $t = t_n$). Within the more general setting of continuous Runge-Kutta methods, this use of defects and relations like (2.21) were discussed earlier, see e.g. [4, 10, 11, 18, 19]. Trivially, multiplying $r(t_{n+1/2})$ by a certain constant multiplies the solution by the same constant (if $e(0) = 0$). This simple property forms the basis for tolerance proportionality which we shall use for attempting control over the global errors (both in the classical and novel approach).

2.3 The example integration formulas

The adjoint approach discussed in Section 3 heavily relies on the small sample statistical method whose major challenge lies in problems of a large dimension. For the comparison between the adjoint and classical approach we will therefore use stiff, semi-discrete diffusion-reaction problems.[2)] Consequently, as example integrator for generating the approximations w_n we have chosen an A-stable scheme, the 3rd-order Runge-Kutta-Rosenbrock scheme ROS3P [16]. To save space we refer to [15, 16] for details.

The implemented step size strategy by which the time grid (1.2) is generated with ROS3P is standard, except that different from [16] the defect $r(t_{n+1/2})$ is used. So for local control we work with

$$Est = (I - \gamma\tau_n A)^{-1} r(t_{n+1/2})\,, \qquad A = F'(t_n, w_n)\,, \tag{2.23}$$

where γ is a ROS3P coefficient. The common filter $(I - \gamma\tau_n A)^{-1}$ serves to damp spurious stiff components which would otherwise be amplified through the F-evaluations within $r(t_{n+1/2})$.[3)] Note that while the local error is $\mathcal{O}(\tau_n^4)$, this estimate is $\mathcal{O}(\tau_n^3)$ by which we asymptotically have tolerance proportionality [17].

Let $D_n = \|Est\|$ with $\|\cdot\|$ the L_2-norm.[4)] The step is accepted if $D_n \le Tol_n$ where $Tol_n = Tol_A + Tol_R\,\|w_n\|$ with Tol_A and Tol_R given tolerances. Otherwise the step is rejected and redone. In both cases the new step size is determined by the standard rule $\tau_{new} = \min\big(1.5, \max(2/3, 0.9\,r)\big)\tau_n$ where $r = (Tol_n/D_n)^{1/3}$. After each step size change we adjust τ_{new} to $\tau_{n+1} = (T - t_n)/\lfloor(1 + (T - t_n)/\tau_{new})\rfloor$ so as to guarantee to reach the end point T with a step of averaged normal length. The initial step size τ_0 is prescribed and is adjusted similarly.

Simultaneously, equation (2.22) is integrated by means of the implicit midpoint rule

$$e_{n+1} = e_n + \tau_n A\Big(\frac{e_{n+1} + e_n}{2}\Big) + \tau_n r(t_{n+1/2})\,, \qquad A = F'(t_n, w_n)\,, \tag{2.24}$$

implemented in the equivalent form

$$\begin{aligned} \tilde{e}_{n+1} &= 2e_n + \tfrac{1}{2}\tau_n A\tilde{e}_{n+1} + \tau_n r(t_{n+1/2})\,, \\ e_{n+1} &= \tilde{e}_{n+1} - e_n\,. \end{aligned} \tag{2.25}$$

The main additional costs for (2.25) come from an extra decomposition since $\gamma \neq \frac{1}{2}$ (assuming a direct solve). Due to freezing $G(t)$ and $\rho(t)$ at $t = t_n$ as discussed above, the second-order midpoint rule (2.24) is a first-order method when interpreted for solving the first variational equation (2.11). As a result, the associated local error takes the form $C(t_n)\tau_n^2$ + h.o.t. where the leading error constant $C(t_n) = \mathcal{O}(\tau_{max}^3)$ as it is proportional to $\eta(t)$ which itself is $\mathcal{O}(\tau_{\max}^3)$. Consequently, the global error approximations e_n satisfy $e_n = \eta(t_n) + \mathcal{O}(\tau_{max}^4)$.

[2)] Observe that only temporal error behaviour is studied and that spatial errors are not discussed.

[3)] The defect contains a new F-evaluation which in turn contains $F(t_n, w_n)$ and $F(t_{n+1}, w_{n+1})$, indicating that two filter steps would be needed. Our practical experience is that one filter step is sufficient, although we know of at least one hypothetical problem (the Kaps problem given at page 215 of [3]) for which two filter steps are appropriate.

[4)] By L_2-norm we mean the weighted inner product norm, i.e., $\|v\|^2 = v^T v/m$, $v \in \mathbb{R}^m$. This norm will also be used for the adjoint approach.

2.4 The control rule

Suppose ROS3P and (2.25) have delivered a numerical solution w_N and a global error estimate e_N at time $t_N = T$. We then verify whether

$$\|e_N\| \le C_{control}\, Tol_N\,, \qquad Tol_N = Tol_A + Tol_R\,\|w_N\|\,, \tag{2.26}$$

where $C_{control} \approx 1$, typically > 1. If (2.26) holds the true global error is considered small enough relative to the chosen tolerance and w_N is accepted. Otherwise, the computation with ROS3P and (2.25) is redone over $[0, T]$ with the same (small) τ_0 and the adjusted tolerances

$$Tol_A = Tol_A \times fac\,, \quad Tol_R = Tol_R \times fac\,, \qquad fac = Tol_N/\|e_N\|\,. \tag{2.27}$$

The primary aim of global error estimation is to provide an additional check on accuracy. Especially when the problem at hand is ill-conditioned (unstable) this is useful since local control will not detect instability so that ε_N might be substantially larger than the imposed tolerance. The second (control) computation with ROS3P (and (2.25) for an additional check) serves to reduce ε_N to the imposed tolerance level. If all is going well, with (2.26) we thus account on the quality of the global error estimation and with (2.27) on tolerance proportionality, thus expecting that reducing the local error estimates with the factor fac will reduce ε_N by fac [17].

Remark 2.2 It is possible to avoid (2.27) by storing the whole step size sequence (1.2) from the first run over $[0, T]$ and to carry out the second computation on a new step size sequence obtained by dividing all intervals $[t_n, t_{n+1}]$ in $\lceil 1/fac^{1/3} \rceil$ equal subintervals. We then only account on the quality of (2.26) and even would have the possibility to also use global Richardson extrapolation for global error estimation for an additional check. However, we then also give up local control. This renders no problem if all is going well, but it might result in instability which otherwise would have been detected by local control. $\diamond$

3 The adjoint (forward-backward) approach

Like for the classical approach the analysis of the adjoint approach starts from the perturbed system derived in Section 2.1.

3.1 Error representation for scalar derived functions

The error representation formula (2.5) there reveals that an approximation of $\Phi^{-1}(t)\,\Phi(s)$, $0 \le s \le t$, would be desirable to estimate the sensitivity of (1.1) with respect to perturbations $v_0 - w_0$ and $r(t)$. Instead of computing these matrix products, which is in general far too expensive or even unattainable, we first derive error estimates for a scalar derived function. The analysis is based on the adjoint method which has been used successfully to obtain goal–oriented a posteriori error estimation for an adaptive error control for PDEs (see [1, 5, 7]) and ODEs (see [2, 6, 13]).

Let $M\big(w(t)\big)$ be the scalar quantity of interest. Then one has for the error in M

$$\triangle M(t) := M\big(v(t)\big) - M\big(w(t)\big) = \overline{M}(t)\, e(t) \tag{3.1}$$

with row vector

$$\overline{M}(t) = \int_0^1 M'\big(v(t) + (s-1)e(t)\big)\, ds = \int_0^1 M'\big(w(t) + se(t)\big)\, ds\,. \tag{3.2}$$

Hence, using (2.5),

$$\triangle M(t) = \overline{M}(t)\Phi^{-1}(t)\Phi(0)e(0) + \overline{M}(t)\Phi^{-1}(t) \int_0^t \Phi(s)r(s)\, ds\,. \tag{3.3}$$

Solving backward in time the adjoint equation

$$\phi'(s) = -A^T(s)\phi(s), \quad \phi(t) = \overline{M}^T(t), \quad 0 \le s < t, \tag{3.4}$$

and taking into account that $\Phi^{-T}(s)$ is the fundamental matrix of this equation if (and only if) $\Phi(t)$ is the fundamental matrix of (2.3) defined in (2.5), one gets

$$\phi^T(s) = \overline{M}(t)\Phi^{-1}(t)\Phi(s)\,. \tag{3.5}$$

Thus

$$\triangle M(t) = \phi^T(0)e(0) + \int_0^t \phi^T(s)r(s)\, ds, \quad 0 \le t \le T. \tag{3.6}$$

Formula (3.6) serves as the fundamental relation to derive a strategy for global error estimation. The adjoint solution value $\phi^T(0)$ measures sensitivity of $M\big(w(t)\big)$ with respect to $e(0)$. Likewise, the integral term measures sensitivity with respect to the defects.

Remark 3.1 In principle one may consider quantities $M\big(w(t)\big) = \xi_i^T w(t),\ i = 1, \ldots, m$, where ξ_i is the i-th unit vector in $\mathbb{R}^m$. Then $\triangle M(t) = \xi_i^T e(t) = e^{(i)}(t)$, the i-th component of the error vector $e(t)$. Thus $\overline{M}^T(t) = \xi_i$. Denoting by ψ_i the solution of (3.4) with ξ_i as initial value, one gets from (3.6)

$$\xi_i^T\, e(t) = e^{(i)}(t) = \psi_i^T(0)e(0) + \int_0^t \psi_i^T(s)r(s)\, ds\,. \tag{3.7}$$

By this way all components of the error vector $e(t)$ could be computed at the price of solving the adjoint equation m times. However, if $m \gg 1$ one would have to solve a tremendous number of adjoint systems making this method impractical. The best choice would be $M(w(t)) = e^T(t)w(t)/(m\|e(t)\|)$ [2]. Then $\triangle M(t) = \|e(t)\|$ directly from (3.1), but one does not have $e(t)$. The choice of appropriate initial conditions for the adjoint system is the main challenge for the adjoint approach [2, 7]. $\diamond$

Remark 3.2 To set up the adjoint equation (3.4) we have to replace $A^T(s)$ by a suitable approximation in the neighborhood of $v(s)$. Thus the adjoint equation depends on the solution of the original ODE. A first possibility is to store the forward solution for every time step to determine the adjoint equation. Alternatively, the solution is stored at only a few selected times $0 = T_0 < T_1 < \cdots < T_K = T$. As the adjoint equation solver marches backwards in time from T_k to T_{k-1}, one recomputes the solution over that time interval using the previously stored solution T_{k-1} as initial value. This approach reduces the storage requirements at the price of a second forward solve. The need to make the forward equation

available to the adjoint equation is clearly a drawback of the adjoint approach. Solution storage is of course not needed for linear systems (1.1) of type $w' = A(t)w + g(t)$. On the other hand, in all cases all defects must be stored. Another drawback of the adjoint approach is that it is defined for single output times t at which estimation is wanted. In other words, if estimation is wanted at multiple times t, the adjoint solution must be computed apart for each value of t. $\diamond$

3.2 Global error estimation

We have implemented global error estimation applying ROS3P to solve (1.1) as described in Section 2.3 and using equation (3.6) and the small sample statistical method proposed by Cao and Petzold for BDF methods ([2], see also [14] for more details). The main idea is to replace the vectors ξ_i in (3.7) by a small number of orthonormal vectors $z_1, z_2, \ldots, z_k$ which are selected uniformly and randomly from the unit sphere S_{m-1} in m dimensions. Instead of computing accurate error estimates of global errors, we try to approximate them with a high probability.

Let $\eta_i(t) = |z_i^T\, e(t)|$. Then an estimate for $\|e(t)\|$ is given by

$$\|e(t)\| \approx g_k(t) = \frac{E_k}{E_m} \sqrt{\frac{1}{m}\left(\eta_1^2(t) + \eta_2^2(t) + \cdots + \eta_k^2(t)\right)}\,, \tag{3.8}$$

where $E_1 = 1$, $E_2 = 2/\pi$, and for $n > 2$

$$E_n = \begin{cases} \dfrac{1 \cdot 3 \cdot 5 \cdots (n-2)}{2 \cdot 4 \cdot 6 \cdots (n-1)} & \text{for } n \text{ odd,} \\[2ex] \dfrac{2}{\pi} \cdot \dfrac{2 \cdot 4 \cdot 6 \cdots (n-2)}{1 \cdot 3 \cdot 5 \cdots (n-1)} & \text{for } n \text{ even.} \end{cases} \tag{3.9}$$

E_n can be estimated by $\sqrt{2/(\pi(n-1/2))}$. The "$\approx$" in (3.8) has to be understood in the sense of probability. More precisely, the expected value of the random variable $g_k(t)$ is given by $E(g_k(t)) = \|e(t)\|$ ([14], Theorem 3.1). For $k = m$ the vectors $z_1, \ldots, z_k$ form an orthonormal basis in $\mathbb{R}^m$ and hence from the definition in (3.8) we get for $k = m$ directly the identity $g_m(t) = \|e(t)\|$.

The important question now is, what is the probability that the estimator $g_k(t)$ provides upper and lower bounds for $\|e(t)\|$? Let $c > 1$ be a given factor. Then one has for two and three random vectors [14]

$$P\left(\frac{g_2(t)}{c} \le \|e(t)\| \le c\, g_2(t)\right) \;\approx\; 1 - \frac{\pi}{4c^2}\,, \tag{3.10}$$

$$P\left(\frac{g_3(t)}{c} \le \|e(t)\| \le c\, g_3(t)\right) \;\approx\; 1 - \frac{32}{3\pi^2\, c^3}\,. \tag{3.11}$$

To achieve 99% probability of accuracy, for example, one can use $k = 2$ for $c = 10$ and $k = 3$ for $c = 5$. In [2] it is pointed out that in practice usually at most two or three random vectors are sufficient, although without giving numerical evidence because in their numerical experiments $k = m$. Hence the small sample statistical method was not tested in [2].

Remark 3.3 To generate k independent vectors $z_1, \ldots, z_k$ uniformly and randomly on S_{m-1} one can use the following procedure: let $\lambda^{(1)}, \ldots, \lambda^{(m)}$ be normally distributed independent random variables with mean zero and variance one. Then the vector $\lambda/\|\lambda\|$ is uniformly and randomly distributed over S_{m-1} where $\lambda = (\lambda^{(1)}, \ldots, \lambda^{(m)})^T$ [22]. $\diamond$

The small sample statistical method described above thus will be used to estimate $\|e(T)\|$. We first select random vectors $z_1, \ldots, z_k$ from S_{m-1} and construct an orthonormal basis for their span by using a Gram–Schmidt procedure or a QR decomposition. Then we solve the corresponding adjoint equations

$$\phi_i'(s) = -A^T(s)\,\phi_i(s)\,, \quad \phi_i(T) = z_i\,, \tag{3.12}$$

for $\phi_i, i = 1, \ldots, k$. From (3.1) and (3.6) with $\triangle M(T) = z_i^T e(T)$ we have for each ϕ_i the identity

$$z_i^T\, e(T) = \phi_i^T(0)e(0) + \int_0^T \phi_i^T(s)r(s)\,ds. \tag{3.13}$$

In the following we assume $e(0) = 0$. Using (3.8) and recalling $\eta_i(T) = |z_i^T\, e(T)|$ we get

$$g_k(T) = \frac{E_k}{E_m}\,\left(\frac{1}{m}\sum_{i=1}^{k}\left(\int_0^T\, \phi_i^T(s)r(s)\,ds\right)^2\right)^{1/2}. \tag{3.14}$$

The integral terms must be approximated (recall that with $k = m$ and an exact computation of these integrals we have $\|e(T)\| = g_m(T)$). We integrate (3.12) using the second-order implicit midpoint rule on the same grid as selected to solve (1.1) by means of ROS3P, but now backward in time starting from z_i

$$\begin{array}{rcl} \tilde{\phi}_{i,n} & = & 2\,\phi_{i,n+1} + \frac{\tau_n}{2}\,A^T\,\tilde{\phi}_{i,n}\,, \quad A^T = \left(F'\left(t_{n+1/2}, \frac{w_n + w_{n+1}}{2}\right)\right)^T, \\ \phi_{i,n} & = & \tilde{\phi}_{i,n} - \phi_{i,n+1}\,, \quad n = N-1, \ldots, 0. \end{array} \tag{3.15}$$

Note that for larger problems one cannot store all Jacobians A from the forward integration, so one has to recompute them. The practical need for recomputing Jacobians is another drawback of the adjoint approach. Further note that like in the forward approach A^T is an accurate approximation to the integrated Jacobian $A(t)$ defined in (2.4) as long as the global error $e(t)$ is sufficiently small. The adjoint problems are not coupled and hence can be solved in parallel. Also recomputing Jacobians for use in the backward in time midpoint rule is needed only once, that is, once for all adjoint problems.

To approximate the integrals in (3.14) we use the 1-point, second-order Gaussian formula for each integration interval to obtain

$$g_k(T) \approx \frac{E_k}{E_m}\,\left(\frac{1}{m}\sum_{i=1}^{k}\left(\sum_{n=0}^{N-1}\,\tau_n\,\frac{\phi_{i,n}^T + \phi_{i,n+1}^T}{2}\,r(t_{n+1/2})\right)^2\right)^{1/2}, \tag{3.16}$$

where the residual $r(t_{n+1/2})$ are computed in exactly the same way as within the classical approach described in Section 2.2 and Section 2.3, i.e., we take $r(t_{n+1/2}) = -\frac{2}{3}d(t_{n+1/2})$.

3.3 The control rule

The possible need for a second forward computation with ROS3P is decided on the same control rule as in the classical approach which we described in Section 2.4. Hence also for the adjoint approach we rely on tolerance proportionality for the global error control. If a second forward computation is decided, then for an additional error check we apply (3.15) once again for the already chosen random vectors z_i taken as initial values, similar as the additional error check in the classical approach.

Remark 3.4 The conditioning of system (1.1) with respect to small perturbations $r(t)$ can be estimated using equation (3.14). There holds

$$\|e(T)\| \approx |g_k(T)| \leq K_T \cdot \max_{0\leq s\leq T} \|r(s)\| \tag{3.17}$$

with the condition number

$$K_T = K_T(\phi_1, \ldots, \phi_k) = \frac{E_k\,\sqrt{m}}{E_m} \left(\sum_{i=1}^{k} \left(\int_0^T \|\phi_i(s)\|\, ds \right)^2 \right)^{1/2} . \tag{3.18}$$

If K_T is small, a well designed defect based local error control will work out well. But if K_T is large, one could end up with a global error much larger than the imposed local tolerance. As proposed in [5] for parabolic problems, and discussed in [2] for BDF methods, one could tighten the local tolerance for the second ROS3P run within (2.27) through $fac = 1/K_T$. However, when taking norms in (3.17), favorable effects of error cancellation and nearly zero defect values are completely ignored and therefore the new local tolerance can be extremely pessimistic. Indeed, this is observed for the example system of Section 4.4. Its condition number computed from (3.18) with $k = 2$, while using the 1-point, second-order Gaussian formula, is huge, being about 10^{20}. On the other hand, the integral terms in (3.14) are of moderate size due to cancellation and many zero entries in the defect function $r(s)$. So for the example system of Section 4.4, the use of the condition number K_T for global error control is impossible. In fact, K_T is of the same size as the a priori condition number $K_{\mu_2} = (e^{\omega T} - 1)/\omega$, based on the logarithmic matrix norm μ_2, given in (2.7). Inserting the accurate bound $\omega = 100$, which in this case is just the largest eigenvalue of the Jacobian matrix due to symmetry [12], yields $K_{\mu_2} \approx 5.2\,10^{19}$. ◇

4 Numerical illustrations

Numerical results are given for (i) a 2-dimensional unstable nonstiff example problem from [2], (ii) the 3-dimensional stable stiff Robertson chemical kinetics problem [8], (iii) a 100-dimensional unstable semi-discrete diffusion-reaction problem (an often used test problem from combustion theory [12]), and (iv) a 400-dimensional unstable semi-discrete diffusion-reaction problem (from pattern formation and often called the bi-stable Allen-Cahn problem). For the semi-discrete problems spatial errors are left out of consideration, i.e., we compare to a highly accurate ODE reference solution.

All four problems are solved with $Tol_A = Tol_R = Tol$ for the four tolerance values $Tol = 10^{-l}, l = 3, 4, 5, 6$ using one and the same initial step size $\tau_0 = 10^{-5}$. If after the first attempt (2.26) is violated with $C_{control} = 1$ a second attempt is carried out with the

same τ_0 and the automatically adjusted new value for Tol through (2.27). Needless to say that $C_{control} = 1$ is too stringent. We use it here only for the sake of illustrating the good performance of the global control rule (2.27). It will be clear from the tables of results whether a second attempt was necessary.

The tables of results contain the following quantities, $Tol_N = Tol\,(1+\|w_N\|)$ from (2.26), for the classical approach the ratio $\|\varepsilon_N\|/\|e_N\|$ of the true global error and the estimated global error, for the adjoint approach the ratio $\|\varepsilon_N\|/g_k(T)$ of the true global error and the estimated global error defined by (3.16), and for both approaches the ratio $\|\varepsilon_N\|/Tol_N$. The first and second ratio serve to illustrate the quality of the estimation, while the third does this for the control. In addition, numbers of accepted and rejected ROS3P steps are given.

Finally, for the small sample statistical initialization used in the adjoint method, $k = m$ orthonormal random vectors were used for the two small sized problems, whereas for the other two much larger problems we used only two random vectors. So only for the two larger problems the small sample statistical initialization was tested and for the two small sized problems the classical and adjoint approach should give identical results, except for minor implementation differences. Also observe that the randomness of the initialization will lead to differences in the results when computations are repeated, although minor ones.

4.1 A low-dimensional nonstiff ODE system

The first test problem is the 2-dimensional unstable linear system [2]

$$w' = \begin{pmatrix} \frac{1}{2(1+t)} & -2t \\ 2t & \frac{1}{2(1+t)} \end{pmatrix} w, \quad w(0) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad 0 < t \le T = 10\,, \tag{4.1}$$

with increasing oscillatory solution $w_1(t) = \cos(t^2)\sqrt{1+t}$, $w_2(t) = \sin(t^2)\sqrt{1+t}$. Table 4.1 shows the results for the classical approach. Since $k = m = 2$, the adjoint approach results are identical except for negligible implementation differences. The quality of the global error estimates is very high and the second runs based on tolerance proportionality yield perfect control for all tolerances.

Tol	Tol_N	$\|\epsilon_N\|/Tol_N$	$\|\epsilon_N\|/\|e_N\|$	Accept	Reject
10^{-3}	$3.32\,10^{-3}$	8.16	1.02	1031	4
$1.25\,10^{-4}$	$3.34\,10^{-3}$	1.03	1.01	2044	0
10^{-4}	$3.34\,10^{-4}$	8.23	1.01	2201	0
$1.22\,10^{-5}$	$3.34\,10^{-4}$	1.00	1.00	4415	0
10^{-5}	$3.34\,10^{-5}$	8.20	1.00	4719	0
$1.22\,10^{-6}$	$3.35\,10^{-5}$	1.00	1.00	9419	0
10^{-6}	$3.35\,10^{-6}$	8.19	1.00	10146	0
$1.22\,10^{-7}$	$3.35\,10^{-6}$	1.00	1.00	20426	0

Table 4.1: Low-dimensional nonstiff ODE system: classical approach.

4.2 A low-dimensional stiff ODE system

Our second test problem is the well-known Robertson problem from the stiff ODE field [8]

$$\begin{aligned} w_1' &= -0.04\, w_1 + 10^4\, w_2 w_3\,, & w_1(0) &= 1\,, \\ w_2' &= \;\; 0.04\, w_1 - 10^4\, w_2 w_3 - 3.0\, 10^7\, w_2^2\,, & w_2(0) &= 0\,, \qquad 0 < t \le T = 1\,. \\ w_3' &= \;\; 3.0\, 10^7\, w_2^2\,, & w_3(0) &= 0\,, \end{aligned} \tag{4.2}$$

This stiff problem is highly stable resulting in global errors much smaller than imposed local tolerances (the effect of the small initial $\tau_0 = 10^{-5}$ is less strong). So, global error control is redundant here and no control runs were carried out. The global error estimation appears to work very well for this stiff problem. Table 4.2 shows results of the adjoint approach. Since $k = m = 3$, the results of the classical approach again are identical, except for negligible implementation differences. These only exist for the column for the values of $\|\epsilon_N\|/g_3(T)$. The corresponding values $\|\epsilon_N\|/\|e_N\|$ for the classical approach are $(1.07, 1.02, 1.03, 1.04)^T$.

Tol	Tol_N	$\|\epsilon_N\|/Tol_N$	$\|\epsilon_N\|/g_3(T)$	Accept	Reject
10^{-3}	$1.56\,10^{-3}$	$7.39\,10^{-5}$	1.05	29	0
10^{-4}	$1.56\,10^{-4}$	$1.05\,10^{-3}$	0.94	31	0
10^{-5}	$1.56\,10^{-5}$	$8.68\,10^{-3}$	1.01	40	1
10^{-6}	$1.56\,10^{-6}$	$7.64\,10^{-2}$	1.02	62	2

Table 4.2: A low-dimensional stiff ODE system: adjoint approach.

4.3 High-dimensional stiff ODE system I

Our third ODE test system is derived from spatially discretizing the well-known 1D combustion example problem (see e.g. [12], p.434, for the 2D version)

$$u_t = u_{xx} + \frac{1}{4}\,(2-u)\,e^{20(1-1/u)}\,, \quad 0 < x < 1\,, \qquad 0 < t \le T = 0.28\,, \tag{4.3}$$

subjected to the initial condition $u(x,0) = 1$, the zero Neumann boundary condition $u_x = 0$ at $x = 0$ and the Dirichlet boundary condition $u = 1$ at $x = 1$. The spatial discretization is done by second-order central differencing on a uniform hybrid grid (because of the Neumann condition) with mesh width $1/100.5$, resulting in a 100-dimensional ODE system $w' = F(w)$. We have chosen this system because it requires variable step sizes and it is unstable. The instability emanates from the reaction term whose derivative ranges between approximately $+1000$ and -5500, see [12] for details. Furthermore, the 100-dimensional ODE system poses a challenging test for the small sample statistical method using $k = 2$ only.

For the classical approach Table 4.3 reveals an excellent estimation of the global error and likewise a high quality of the control, for all tolerances. Observe that for $Tol = 10^{-6}$ the second control run is redundant. Regarding the adjoint approach we emphasize the high quality of the small sample statistical method. From inequality (3.10) one would expect the ratios $\|\varepsilon_N\|/g_2(T)$ to lie asymptotically between 0.1 and 10.0 with 99% probability. Table 4.4 however shows ratio values ranging between 0.72 and 1.38. Taking k larger with this large dimension m is no option. For example, doubling k to 4 does not lead to a notable improvement, yet increases the costs for the global error estimation. Noting the small

deviations from 1.0 of all listed ratios, especially after the control run, both approaches perform excellent with respect to estimation and control.

Tol	Tol_N	$\|\epsilon_N\|/Tol_N$	$\|\epsilon_N\|/\|e_N\|$	Accept	Reject
10^{-3}	$2.83\,10^{-3}$	2.56	1.25	529	33
$4.91\,10^{-4}$	$2.84\,10^{-3}$	1.03	1.20	680	32
10^{-4}	$2.84\,10^{-4}$	2.64	1.13	1183	18
$4.30\,10^{-5}$	$2.84\,10^{-4}$	1.11	1.09	1586	11
10^{-5}	$2.84\,10^{-5}$	2.09	1.05	2622	5
$5.02\,10^{-6}$	$2.84\,10^{-5}$	0.85	1.03	3318	3
10^{-6}	$2.84\,10^{-6}$	0.91	1.00	5736	3

Table 4.3: High dimensional stiff ODE system I: classical approach.

Tol	Tol_N	$\|\epsilon_N\|/Tol_N$	$\|\epsilon_N\|/g_2(T)$	Accept	Reject
10^{-3}	$2.83\,10^{-3}$	2.56	1.35	529	33
$5.26\,10^{-4}$	$2.84\,10^{-3}$	1.06	1.38	664	32
10^{-4}	$2.84\,10^{-4}$	2.64	0.92	1183	18
$3.47\,10^{-5}$	$2.84\,10^{-4}$	0.84	0.92	1708	8
10^{-5}	$2.84\,10^{-5}$	2.09	0.76	2622	5
$3.66\,10^{-6}$	$2.84\,10^{-5}$	0.57	0.76	3695	4
10^{-6}	$2.84\,10^{-6}$	0.91	0.72	5736	3
$7.93\,10^{-7}$	$2.84\,10^{-6}$	0.65	0.72	6204	3

Table 4.4: High dimensional stiff ODE system I: adjoint approach.

4.4 High-dimensional stiff ODE system II

Similar to the third, the fourth test problem was chosen because it is unstable again challenging global error control. It is derived from spatially discretizing the following version of the bi-stable Allen-Cahn equation

$$u_t = 10^{-2}\, u_{xx} + 100u\,(1-u^2)\,, \quad 0 < x < 2.5\,, \qquad 0 < t \le T = 0.5\,, \tag{4.4}$$

with the initial function and Dirichlet boundary values taken from the exact wave front solution $u(x,t) = (1+e^{\lambda\,(x-\alpha\,t)})^{-1}$, $\lambda = 0.5\,\sqrt{2}$, $\alpha = 1.5\,\sqrt{2}$. Uniform second-order central discretization in space yields our ODE test system, now with $m = 400$ components. The instability emanates from the unstable stationary state $u = 0$. Further, since $m = 400$ and $k = 2$, also this problem poses an even more challenging test for the small sample statistical method.

Table (4.5) reveals again a high quality of the global error estimation for the classical approach and also the control process works very well. The ratios for $\|\epsilon_N\|/Tol_N$ lie between 0.71 and 0.93, after the control run. Observe that for $Tol = 10^{-5},\ 10^{-6}$ the control runs are redundant. In actual practice, taking the control parameter $C_{control} > 1$ in (2.26), this would also hold for the other tolerances. For the adjoint approach, the ratios $\|\epsilon_N\|/g_2(T)$ given in Table (4.6) range from 0.53 to 2.56, which is by a factor 4 better than one would

expect from inequality (3.10). The adjoint approach does not require a second run for $Tol = 10^{-4},\ 10^{-6}$ and for the other cases the control is also very efficient (only a factor 0.45 to 0.78 off the imposed tolerance, after the control run).

Tol	Tol_N	$\|\epsilon_N\|/Tol_N$	$\|\epsilon_N\|/\|e_N\|$	Accept	Reject
10^{-3}	$1.65\,10^{-3}$	1.29	0.77	373	0
$6.01\,10^{-4}$	$1.65\,10^{-3}$	0.71	0.83	446	0
10^{-4}	$1.65\,10^{-4}$	0.95	0.93	833	0
$9.84\,10^{-5}$	$1.65\,10^{-4}$	0.93	0.93	838	0
10^{-5}	$1.65\,10^{-5}$	0.82	0.97	1835	0
10^{-6}	$1.65\,10^{-6}$	0.76	0.98	3998	0

Table 4.5: High-dimensional stiff ODE system II: classical approach.

Tol	Tol_N	$\|\epsilon_N\|/Tol_N$	$\|\epsilon_N\|/g_2(T)$	Accept	Reject
10^{-3}	$1.65\,10^{-3}$	1.29	0.53	373	0
$4.09\,10^{-4}$	$1.65\,10^{-3}$	0.45	0.57	510	0
10^{-4}	$1.65\,10^{-4}$	0.95	1.40	833	0
10^{-5}	$1.65\,10^{-5}$	0.82	0.78	1835	0
$9.45\,10^{-6}$	$1.65\,10^{-5}$	0.78	0.78	1870	0
10^{-6}	$1.65\,10^{-6}$	0.76	2.56	3998	0

Table 4.6: High-dimensional stiff ODE system II: adjoint approach.

5 Summary and main conclusions

Inspired by [2] and related earlier literature, e.g. [1, 5, 6, 7, 13], we have discussed and compared classical global error estimation based on the first variational equation to a recent more novel approach based on the adjoint equation. The common starting point for both approaches is the perturbed equation with the residual or defect function defined by piecewise, cubic Hermite interpolation. As a base integrator for the comparison we have chosen the third-order, A-stable Runge-Kutta-Rosenbrock method ROS3P. We have also implemented global error control, for which we have used the property of tolerance proportionality for both the classical and the adjoint approach.

On the basis of the four example problems, ranging from nonstiff to stiff, the computational effort of each of the two approaches, and the insight from the analysis, we have come to three main conclusions. (i) The classical approach is remarkably reliable, both with respect to estimation and control. Although well known in the numerical ODE literature, it seems that the virtue of this approach has been insufficiently brought forward since as yet this form of global error estimation and control is far less popular than the commonly used local techniques. (ii) Most notable for the adjoint approach is the excellent performance of the small sample statistical method which forms the heart of the method. We have applied it successfully using only $k = 2$ random vectors for dimensions $m = 100, 400$ (in [2] it was not tested since there $k = m \leq 2$). With $k = 2$ the computational costs are only marginally

higher than those of the classical approach. (iii) The main disadvantage of the adjoint approach is the need to either store the whole approximation sequence $(w_n; 0 \leq 1 \leq N)$ or to store part of it and to carry out a second forward computation. When using Jacobians, as is the case for ROS3P, storage becomes truly a handicap with large problems calling for Jacobian reevaluations and hence additional CPU costs. The classical approach does not suffer from this storage handicap, and because it is also clearly competitive to the adjoint approach with respect to the quality of estimation and control, we consider it more attractive, more efficient, and significantly more practical for large dimensional ODE systems (1.1).

References

[1] R. Becker, R. Rannacher (2001), *An optimal control approach to a posteriori error estimation in finite element methods*, Acta Numerica 10, pp. 1-102.

[2] Y. Cao, L. Petzold (2004), *A posteriori error estimation and global error control for ordinary differential equations by the adjoint method*, SIAM J. Sci. Comput. 26, pp. 359-374.

[3] K. Dekker, J.G. Verwer (1984), *Stability of Runge-Kutta Methods for Stiff Nonlinear Differential Equations*, North-Holland, Elsevier, Amsterdam.

[4] W.H. Enright (2000), *Continuous numerical methods for ODEs with defect control*, J. Comp. Appl. Math. 125, pp. 159-170.

[5] K. Eriksson, C. Johnson, A. Logg (2004), *Adaptive computational methods for parabolic problems*, Chapter 24 of Encyclopedia of Computational Mechanics, eds. E. Stein, R. de Borst, and J.R. Hughes, John Wiley & Sons, Ltd.

[6] D. Estep (1995), *A posteriori error bounds and global error control for approximation of ordinary differential equations*, SIAM J. Numer. Anal. 32, pp. 1-48.

[7] D. Estep, M. Larsson, R. Williams (2000), *Estimating the error of numerical solutions of systems of reaction–diffusion equations*, Mem. Amer. Math. Soc. 146.

[8] E. Hairer, S.P. Nørsett, G. Wanner (1993), *Solving Ordinary Differential Equations I - Nonstiff Problems*, Springer Series in Computational Mathematics, Vol. 8, Second edition, Springer, Berlin.

[9] P. Henrici (1962), *Discrete Variable Methods in Ordinary Differential Equations*, John Wiley & Sons, New York.

[10] D.J. Higham (1989), *Robust defect control with Runge-Kutta schemes*, SIAM J. Numer. Anal. 26, pp. 1175-1183.

[11] D.J. Higham (1991), *Runge-Kutta defect control using Hermite-Birkhoff interpolation*, SIAM J. Sci. Stat. Comput. 12, pp. 991-999.

[12] W. Hundsdorfer, J.G. Verwer (2003), *Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations*, Springer Series in Computational Mathematics, Vol. 33, Springer, Berlin.

[13] C. Johnson (1988), *Error estimates and adaptive time–step control for a class of one–step methods for stiff ordinary differential equations*, SIAM J. Numer. Anal. 25, pp. 908-926.

[14] C.S. Kenney, A.J. Laub (1994), *Small–sample statistical condition estimates for general matrix functions*, SIAM J. Sci. Comput. 15, pp. 36-61.

[15] J. Lang (2000), *Adaptive Multilevel Solution of Nonlinear Parabolic PDE Systems. Theory, Algorithm and Applications*, Lecture Notes in Computational Science and Engineering, Vol. 16, Springer.

[16] J. Lang, J.G. Verwer (2001), *ROS3P – An accurate third-order Rosenbrock solver designed for parabolic problems*, BIT 41, pp. 731-738.

[17] L.F. Shampine (1994), *Numerical Solution of Ordinary Differential Equations*, Chapman & Hall, New York.

[18] L.F. Shampine (2005), *Error estimation and control for ODEs*, J. of Scientific Computing, to appear.

[19] L.F. Shampine (2005), *Solving ODEs and DDEs with residual control*, Appl. Numer. Math. 52, pp. 113-127.

[20] R.D. Skeel (1986), *Thirteen ways to estimate global error*, Numer. Math. 48, pp. 1-20.

[21] H.J. Stetter (1974), *Economical global error estimation*, in: Stiff Differential Systems, ed. R.A. Willoughby, Plenum Press, New York, London, pp. 245-258.

[22] G. Watson, *Statistics on Spheres*, Wiley, New York, 1983; Vol. 6 in the Univ. of Arkansas Lecture Notes in the Mathematical Science

[23] P.E. Zadunaisky (1976), *On the estimation of errors propagated in the numerical integration of ordinary differential equations*, Numer. Math. 27, pp. 21-39.