



# NEW RUNGE–KUTTA BASED SCHEMES FOR ODES WITH CHEAP GLOBAL ERROR ESTIMATION\* \*

J. MAKAZAGA<sup>1</sup> and A. MURUA<sup>1</sup>

<sup>1</sup>*Konputazio Zientziak eta Adimen Artifiziala, University of the Basque Country (EHU/UPV), Informatika Fakultatea, Paseo Lardizabal, Donostia-San Sebastian 2080, Spain. email: {joxe, ander}@si.ehu.es*

## Abstract.

We present a particular 5th order one-step integrator for ODEs that provides an estimation of the global error. It's based on the class of one-step integrator for ODEs of Murua and Makazaga considered as a generalization of the globally embedded RK methods of Dormand, Gilmore and Prince. The scheme we present cheaply gives useful information on the behavior of the global error. Some numerical experiments show that the estimation of the global error reflects the propagation of the true global error. Moreover we present a new step-size adjustment strategy that takes advantage of the available information about the global error. The new strategy is specially suitable for problems with exponential error growth.

*AMS subject classification (2000):* 65L06, 65L50.

*Key words:* Runge–Kutta methods, global error estimation, variable step-size strategy.

## 1 Introduction.

Very efficient methods exist for the numerical integration of nonstiff ODEs. They are intended to maintain the global error controlled keeping the local error below a prescribed tolerance. However, they do not provide any information about the global error, unless additional computational effort is done.

Interesting work in obtaining schemes that give information about the propagation of the global error has been done by Dormand, Gilmore and Prince [3]. In this paper we present a methodology to construct methods of integration for nonstiff problems based on the general class of schemes presented in [6]. These schemes are a generalization of the methods due to Dormand, Gilmore and Prince.

---

\* Received: August 2002. Revised June 2003. Communicated by Timo Eirola.

\* This work was partially supported by grant Order Foral 328/99 of Gipuzkoako Foru Aldundia.

We consider (nonstiff) ODE systems in autonomous form

$$(1.1) \quad \frac{dy}{dt} = f(y), \quad y \in \mathbb{R}^D, \quad f: \mathbb{R}^D \rightarrow \mathbb{R}^D.$$

We wish to integrate initial value problems of the form (1.1) over an interval  $[t_0, t_{\text{end}}]$  and we want to get, together with the numerical solution, an estimation of the global error.

Recall that the  $h$ -flow of the system is a parametric transformation of phase space  $\phi_h: \mathbb{R}^D \rightarrow \mathbb{R}^D$  such that  $\phi_h(y(t)) = y(t+h)$ , and when we solve numerically the ODE the flow  $\phi_h$  is replaced by a transformation of phase space  $\psi_h: \mathbb{R}^D \rightarrow \mathbb{R}^D$  that approximates it. The local error is defined as  $\delta(y, h) = \psi_h(y) - \phi_h(y)$ , and the method  $\psi_h$  is of order  $p$  if  $\delta(y, h) = O(h^{p+1})$ .

For a given time discretization  $t_0 < t_1 < \dots < t_N = t_{\text{end}}$  with  $h_n = t_n - t_{n-1}$  the integrator computes the numerical solution

$$(1.2) \quad y_n = \psi_{h_n}(y_{n-1}), \quad n = 1, \dots, N$$

as an approximation to

$$y(t_n) = \phi_{h_n}(y(t_{n-1})), \quad n = 1, \dots, N.$$

The global error  $e_n = y_n - y(t_n)$  satisfies that

$$(1.3) \quad \begin{aligned} e_n &= \psi_{h_n}(y_{n-1}) - \phi_{h_n}(y(t_{n-1})) \\ &= (\phi_{h_n}(y(t_{n-1}) + e_{n-1}) - \phi_{h_n}(y(t_{n-1}))) + \delta(y_{n-1}, h_n). \end{aligned}$$

This shows that the global error  $e_n$  can be viewed as the sum of two different errors,

- The local error  $\delta(y_{n-1}, h_n) = \psi_{h_n}(y_{n-1}) - \phi_{h_n}(y_{n-1})$  committed at each step.
- Propagation of the global error at previous step, or equivalently, propagation and accumulation of local errors made at previous steps.

For the family of explicit Runge–Kutta (RK) methods,  $\psi_h$  is defined as

$$(1.4) \quad \psi_h(y) = y + h \sum_{i=1}^s b_i f(Y_i),$$

where for  $i = 1, \dots, s$

$$(1.5) \quad Y_i = y + h \sum_{j=1}^{i-1} a_{i,j} f(Y_j).$$

In this paper we consider a subfamily of the class of one-step integrators for ODEs introduced in [6]. The general class is briefly introduced in Section 2. The conditions on the parameters that should hold for a good method of that family

are considered in Section 3. In Section 4, a particular choice of the parameters of the method is presented. Numerical tests are presented in Section 5 for several initial value problems, where the global error committed by our method is compared with the provided estimation of the global error. In Section 6, a new strategy to adjust the step-size is proposed. This strategy takes advantage of the available information about the global error. Finally, some numerical experiments are presented to test the new strategy.

**2 A general class of schemes.**

Substitution of  $y(t_{n-1}) = y_{n-1} - e_{n-1}$  in (1.3) gives a recurrence for the global error of the form

$$(2.1) \quad e_n = E_{h_n}(y_{n-1}, e_{n-1}),$$

where the mapping  $E_h: \mathbb{R}^{2D} \rightarrow \mathbb{R}^D$  is defined by

$$E_h(y, e) = \psi_h(y) - \phi_h(y - e) = \phi_h(y) - \phi_h(y - e) + \delta(y, h).$$

Obviously, if the exact flow is not available, that mapping  $E_h$  will not be available either. If we could compute a mapping  $\tilde{E}_h$  that somehow approximates the true global mapping  $E_h$  we would be able to estimate  $\tilde{e}_n$  as

$$(2.2) \quad \tilde{e}_n = \tilde{E}_h(y_{n-1}, \tilde{e}_{n-1}), \quad n = 1, \dots, N.$$

Extrapolation would then provide a second approximation  $\bar{y}_n = y_n - \tilde{e}_n$ . Thus, the process of obtaining the approximations  $y_n$  together with the global error estimates  $\tilde{e}_n$  can be alternatively interpreted as obtaining two approximations  $y_n$  and  $\bar{y}_n$ , and then computing the estimated global error as their difference. More specifically, define  $\bar{\psi}_h(y, \bar{y}) = \psi_h(y) - \tilde{E}_h(y, y - \bar{y})$ , and compute

$$(2.3) \quad y_n = \psi_{h_n}(y_{n-1}), \quad \bar{y}_n = \bar{\psi}_{h_n}(y_{n-1}, \bar{y}_{n-1}), \quad \tilde{e}_n = y_n - \bar{y}_n.$$

In particular, *Richardson extrapolation* and the *globally embedded RK methods* due to Dormand, Gilmore, and Prince fit into the format (2.3).

A possible generalization of processes of the form (2.3) is as follows: Take  $y_0 = \bar{y}_0 = y(t_0)$  and compute for  $n = 1, 2, \dots, N$

$$(2.4) \quad y_n = \psi_{h_n}(y_{n-1}, \bar{y}_{n-1}), \quad \bar{y}_n = \bar{\psi}_{h_n}(y_{n-1}, \bar{y}_{n-1}), \quad \tilde{e}_n = y_n - \bar{y}_n.$$

*2.1 Generalized globally embedded RK schemes.*

The schemes proposed in [6] fits in the general format (2.4) where the mappings  $\psi_h, \bar{\psi}_h: \mathbb{R}^{2D} \rightarrow \mathbb{R}^D$  are defined by

$$(2.5) \quad \psi_h(y, \bar{y}) = y + h \sum_{i=1}^{\bar{s}} b_i f(Y_i),$$

$$(2.6) \quad \bar{\psi}_h(y, \bar{y}) = \bar{y} + h \sum_{i=1}^{\bar{s}} \bar{b}_i f(Y_i),$$

where for  $i = 1, \dots, \bar{s}$ ,

$$(2.7) \quad Y_i = \mu_i y + (1 - \mu_i) \bar{y} + h \sum_{j=1}^{i-1} a_{i,j} f(Y_j).$$

The parameters  $b_i, \bar{b}_i, a_{ij}, \mu_i$  have to be chosen in such a way that, when applying (2.4),  $y_n$  and  $\bar{y}_n$  approximate  $y(t_n)$  and  $\tilde{e}_n = y_n - \bar{y}_n$  is an useful estimate to the global error. A scheme of the form (2.5)–(2.7) has been presented in [6]. This scheme works particularly well for integration in constant step-size mode. Nevertheless, efficient software requires variable step-sizes, and the usual strategies for step-size adjustment require a cheap estimate of the local error computed from the available intermediate values  $f(Y_i)$ . However, these intermediate values in general depend on both approximations  $y_{n-1}$  and  $\bar{y}_{n-1}$  and therefore are affected by the size of global error estimate  $\tilde{e}_{n-1}$ . This causes some difficulties, which can easily be avoided (at the expense of losing some parameters of the method) as follows. If the following conditions are required

$$(2.8) \quad \mu_i = 1, \quad i = 1, \dots, s,$$

$$(2.9) \quad b_i = 0, \quad i = s + 1, \dots, \bar{s},$$

then the mapping  $\psi_h$  is made independent of  $\bar{y}$  (actually,  $\psi_h$  becomes the RK map (1.4)–(1.5)) and the required local error estimates can be obtained in the standard way (see [7], pp. 334–340, and [5], p. 167)). In fact, by requiring (2.8)–(2.9) we are forcing our scheme to be of the form (2.3), so that the global error estimation process can be interpreted as replacing the recurrence (2.1) by (2.2), with  $\tilde{E}_h(y, h) := \psi_h(y) - \bar{\psi}_h(y, y - e)$ . In what follows, we will focus in the subclass of the Generalized Globally Embedded RK schemes satisfying (2.8)–(2.9). If in addition to (2.8)–(2.9) the following conditions hold,

$$\bar{b}_i = 0, \quad i = 1, \dots, s,$$

$$\mu_i = 0, \quad i = s + 1, \dots, \bar{s},$$

then the family of globally embedded RK schemes of Dormand, Gilmore and Prince [3] is obtained.

### 3 Conditions on the parameters of the method.

In order to construct a method of the form (2.4) one has to determine appropriate values for the parameters  $b_i, \bar{b}_i, a_{ij}$  and  $\mu_i$ . These parameters must be chosen (given  $s$  and  $\bar{s}$ ) according to the following criteria

- (1) The local error  $\delta(y, h) = \psi_h(y) - \phi_h(y)$  of the RK scheme is as small as possible. Typically, one requires that  $\delta(y, h) = O(h^{p+1})$  for a certain  $p \geq 1$  (i.e., that the method be of order  $p$ ) with reasonably small error constants in  $O(h^{p+1})$ .

- (2) The difference  $\bar{\psi}_h(y + e, y) - \phi_h(y)$  is as small as possible. The smaller this difference, the more similar will be the map  $\tilde{E}_h(y, e)$  to the global error map  $E_h(y, e)$ . Our goal will be obtaining estimates of  $\bar{\psi}_h(y + e, y) - \phi_h(y)$  of the form

$$(3.1) \quad \psi_h(y + e, y) - \phi_h(y) = O(h^{q_0} + h^{q_1} \|e\| + h^{q_2} \|e\|^2 + \dots)$$

with sufficiently high  $q_k$  and small error constants.

The first criterion is standard for explicit RK methods. It is well known how to obtain the conditions on the coefficients of a RK method to achieve a prescribed order using rooted trees [2, 5]. We will refer to these rooted trees as RK-trees. Each RK-tree is associated to one condition in terms of the parameters  $b_i, a_{ij}$  of the method, and the RK method achieve order  $p$  if the conditions corresponding to all RK-trees with  $p$  or fewer number of vertices are satisfied.

As for the second criterion, (3.1) holds with given  $q_k$  ( $k = 0, 1, 2, \dots$ ) if certain conditions on the parameters  $\bar{b}_i, a_{ij}, \mu_i$  hold. There is a nice correspondence between such conditions and the set of rooted trees with vertices of two colors. Table 3.1 shows the correspondence between the first conditions and rooted trees. Hereafter we use the notation  $c_i = \sum a_{ij}$ . The exponent  $q_0$  in (3.1) depends on whether certain conditions written in terms of the parameters  $\bar{b}_i, a_{ij}$  (but not in the parameters  $\mu_i$ ) hold. Such conditions are those associated with RK-trees. This is not surprising, since  $q_0 - 1$  is the order of the RK tree defined by the map  $\bar{\psi}_h(y, y)$ , i.e., by the RK scheme determined by the parameters  $\bar{b}_i, a_{ij}$ . In general, the exponent  $q_k$  in (3.1) depends on the conditions corresponding to rooted trees with  $k$  white vertices.

#### 4 Construction of a method of order 5.

We have constructed a scheme (2.5)–(2.7) satisfying (2.8)–(2.9) based on the well known 5th order explicit RK method of Dormand and Prince [4] (we will refer to it as DOPRI5). More precisely, the parameters  $s, b_j, a_{ij}$  ( $1 \leq i \leq s$ ) are determined in such a way that  $\psi_h(y)$  coincides with one step of DOPRI5. In particular,  $s = 6$ . In addition, we have chosen  $\bar{s} = 10$ . Similarly to what is standard in the construction of (locally) embedded RK schemes, we take  $y_{n-1}$  and  $y_n$  respectively as the first and last stages of the scheme (2.5)–(2.7). This is achieved by taking  $a_{7i} = b_i$  ( $1 \leq i \leq 6$ ),  $\mu_1 = 1$  and  $\mu_7 = 1$ .

The remaining parameters  $a_{ij}$  ( $8 \leq i \leq 10, j < i$ ),  $\mu_8, \mu_9, \mu_{10}$  and  $\bar{b}_i$  ( $1 \leq i \leq 10$ ) have been determined in such a way that the estimate  $\bar{\psi}(y, \bar{y}) - \phi(\bar{y}) = O(h^7 + h^4 \|e\| + h^2 \|e\|^2 + h \|e\|^3)$  holds (i.e.,  $q_0 = 7, q_1 = 4, q_2 = 2, q_3 = 1$  in (3.1)). This leaves us with 10 free parameters, which we try to use to minimize the error constants of the terms corresponding to  $h^7, h^4 \|e\|$  and  $h^2 \|e\|^2$ . More precisely, we have tried to minimize numerically in the mean square sense the conditions (obtained as in Table 3.1) that would be required for the estimate (3.1) to hold

Table 3.1: Correspondence between trees, conditions imposed on the coefficients, and estimates of error























Tree	Condition	Estimate of $\overline{\psi}_h(y, y + e) - \phi_h(y)$
	$\sum_i \bar{b}_i = 1$	$O(h^3 + h^2\ e\  + h\ e\ ^2)$
	$\sum_i \bar{b}_i c_i = 1$	
	$\sum_i \bar{b}_i \mu_i = 0$	
	$\sum_{i,j} \bar{b}_i a_{ij} c_j = \frac{1}{6}$	$O(h^4 + h^2\ e\  + h\ e\ ^2)$
	$\sum_i \bar{b}_i c_i^2 = \frac{1}{3}$	
	$\sum_{i,j} \bar{b}_i a_{ij} \mu_j = 0$	$O(h^4 + h^3\ e\  + h^2\ e\ ^2 + h\ e\ ^3)$
	$\sum_{i,j} \bar{b}_i c_i \mu_i = 0$	
	$\sum_i \bar{b}_i \mu_i^2 = 0$	
	$\sum_i \bar{b}_i \mu_i^2 = 0$	
	$\sum_{i,j,k} \bar{b}_i a_{ij} a_{jk} c_k = \frac{1}{24}$	$O(h^5 + h^3\ e\  + h^2\ e\ ^2 + h\ e\ ^3)$
	$\sum_{i,j} \bar{b}_i a_{ij} c_j^2 = \frac{1}{12}$	
	$\sum_{i,j} \bar{b}_i c_i a_{ij} c_j = \frac{1}{8}$	
	$\sum_i \bar{b}_i c_i^3 = \frac{1}{4}$	
	$\sum_{i,j,k} \bar{b}_i a_{ij} a_{jk} \mu_k = 0$	$O(h^5 + h^4\ e\  + h^2\ e\ ^2 + h\ e\ ^3)$
	$\sum_{i,j} \bar{b}_i a_{ij} c_j \mu_j = 0$	
	$\sum_{i,j} \bar{b}_i c_i a_{ij} \mu_j = 0$	
	$\sum_{i,j} \bar{b}_i \mu_i a_{ij} c_j = 0$	
	$\sum_{i,j} \bar{b}_i \mu_i c_j^2 = 0$	
	$\sum_{i,j} \bar{b}_i a_{ij} \mu_j^2 = 0$	$O(h^5 + h^4\ e\  + h^3\ e\ ^2 + h\ e\ ^3)$
	$\sum_{i,j} \bar{b}_i \mu_i a_{ij} \mu_j = 0$	
	$\sum_i \bar{b}_i \mu_i^2 c_i = 0$	
	$\sum_{i,j} \bar{b}_i \mu_i^3 = 0$	$O(h^5 + h^4\ e\  + h^3\ e\ ^2 + h^2\ e\ ^3 + h\ e\ ^4)$

Table 4.1: Remaining parameters of the method based on DOPRI5

$i$	$1 - \mu_i$	$c_i$	$a_{8,i}$	$a_{9,i}$	$a_{10,i}$	$\bar{b}_i$
1	0	0	$\frac{26251126}{75292183}$	$-\frac{126276029}{115017392}$	$\frac{89178409}{82486612}$	$\frac{56696811}{789712427}$
2	0	$\frac{1}{5}$	$-\frac{30511879}{68834945}$	$\frac{153409379}{49308629}$	$-\frac{275044175}{99029299}$	0
3	0	$\frac{3}{10}$	$\frac{11490887}{155205387}$	$-\frac{107711621}{48274693}$	$\frac{115406143}{68971088}$	$-\frac{47431484}{279691831}$
4	0	$\frac{4}{5}$	$\frac{700737845}{174891007}$	$-\frac{675136779}{64711289}$	$\frac{140298385}{24130572}$	$\frac{72791025}{357831874}$
5	0	$\frac{8}{9}$	$-\frac{5336}{941}$	$\frac{559269939}{36928210}$	$-\frac{344040692}{42025591}$	$\frac{17490085}{349505178}$
6	0	1	$\frac{5735}{1214}$	$-\frac{669687859}{52442748}$	$\frac{121333564}{17575013}$	$-\frac{66245097}{563676842}$
7	0	1	$-\frac{2507}{898}$	$\frac{193952703}{25738526}$	$-\frac{190380249}{47005513}$	$-\frac{24}{611}$
8	$\frac{140719960}{143529893}$	$\frac{204}{823}$	0	$\frac{169021117}{130072535}$	$-\frac{12078143}{165601005}$	$\frac{40757463}{82884629}$
9	$\frac{941}{896}$	$\frac{579}{1036}$	0	0	$\frac{56747365}{92317949}$	$\frac{33159666}{111811519}$
10	$\frac{92493035}{95359057}$	1	0	0	0	$\frac{42422453}{199331202}$

with  $q_0 = 8$ ,  $q_1 = 5$ ,  $q_2 = 3$ ,  $q_3 = 2$ . After an extensive numerical search, we have chosen a particular set of values for the free parameters, that determines the parameters  $a_{ij}$  ( $8 \leq i \leq 10$ ,  $j < i$ ),  $\mu_8$ ,  $\mu_9$ ,  $\mu_{10}$  and  $\bar{b}_i$  ( $1 \leq i \leq 10$ ) of the proposed method. They are given in Table 4.1, rationalized within an accuracy of  $10^{-20}$ .

Modern numerical integration codes usually require the availability of reliable interpolants for dense output (i.e. for off-grid points). Our fifth-order method can be implemented with the dense output of the underlying standard Runge–Kutta scheme ([5], pp. 191–192) without being affected by the global error estimate. It is also possible constructing a different dense output formula that takes advantage of the additional stage values  $Y_i$ ,  $i = s + 1, \dots, \bar{s}$ , which would in principle give a more accurate interpolant, at the expense of being dependent of the global error estimate.

## 5 Numerical experiments.

The behavior of the estimated global error has been tested with some numerical experiments. The aim of the experiments is to check whether our method gives useful global error estimates when implemented in variable step-size mode. We have considered three initial value problems:

- (1) The first one has been taken from [5]. It is a 4th dimensional initial value problem, referred as *Arenstorf*. Its initial values correspond to a periodic solution of the restricted 3-body problem.

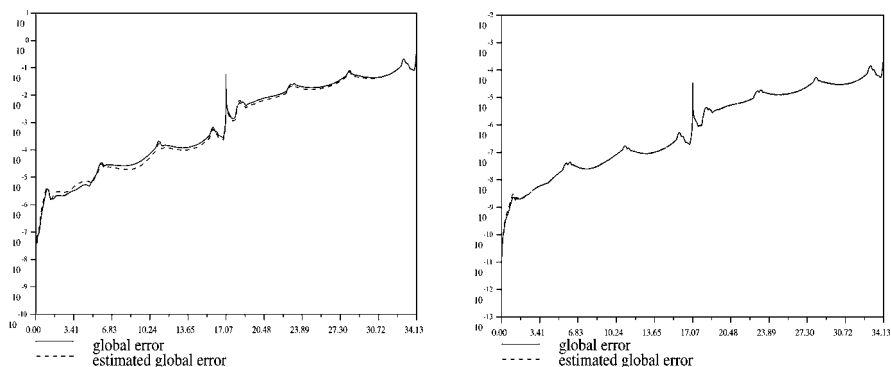


Figure 5.1: The *Arenstorf* problem solved with a tolerance of  $10^{-6}$  (left, 309 steps) and  $10^{-9}$  (right, 1268 steps).

- (2) The second one, referred as *Pleiades*, is part of the test set for Initial Value Problems IVPTestSet [8]. It is a 28th dimensional problem.
- (3) We have taken the third problem from [7]. The system is of dimension  $D = 1$ , and the initial value problem is defined as

$$y' = \cos(t)y, \quad y(0) = 1.$$

Its solution is  $y(t) = e^{\sin(t)}$ , a periodic function with period  $2\pi$ . We refer to this problem as *expsin*.

For each of the problems we display two plots, one corresponds to a stringent tolerance and the other one to a lax tolerance. The plots of Figure 5.1 correspond to the *Arenstorf* problem. It has been integrated over two periods. The two plots show time versus infinity norm of the error in double logarithmic scale. Both plots compare the estimated global error (dashed line) with the exact global error (continuous line). The first one has been integrated with a tolerance of  $10^{-6}$  while in the second one the more stringent tolerance of  $10^{-9}$  is used. The results show that, for the considered initial value problems, the estimated global error behaves in a similar way to the true global error.

The results of the integration of the *Pleiades* problem can be seen in Figure 5.2. Again the reader can see two plots, each one refers to a different tolerance ( $10^{-4}$  and  $10^{-9}$ ) and in both plots it can be seen that the estimated global error (dashed line) behaves as a small perturbation of the true global error (continuous line). In both problems the more stringent tolerance the more accurate estimation of the global error is obtained.

Figure 5.3 shows the results for the third problem. Again we have two plots, the first one corresponds to an absolute tolerance of  $10^{-4}$  and the other one to  $10^{-9}$ . This problem has a greater difference within the true global error and the estimated global error, but the behaviors in both plots are similar.



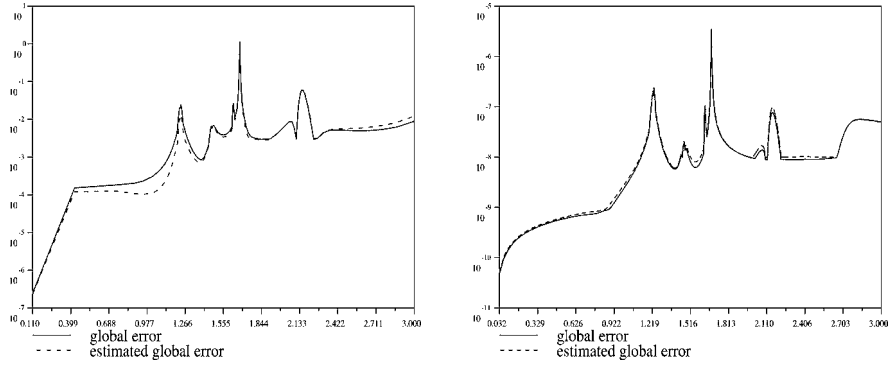


Figure 5.2: The *Pleiades* problem solved with a tolerance of  $10^{-4}$  (left, 182 steps) and  $10^{-9}$  (right, 1603 steps).

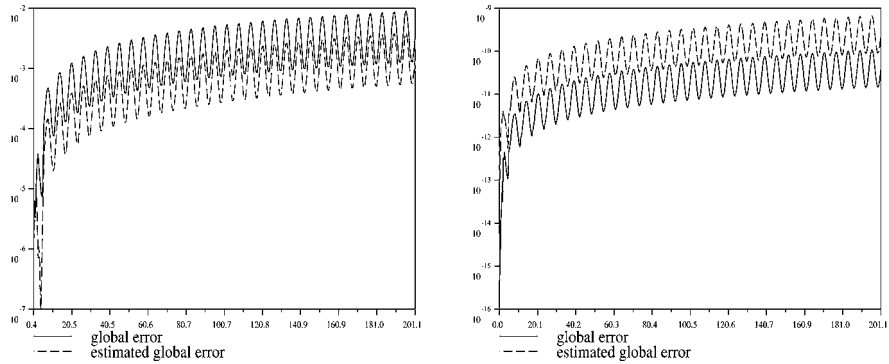


Figure 5.3: The *Expsin* problem solved with a tolerance of  $10^{-4}$  (left, 416 steps) and  $10^{-9}$  (right, 7467 steps).

**6 A new variable step-size strategy.**

We want to explore the possibility of using available information about the propagation of the global error provided by our integrator to improve the efficiency of the step-size adjustment strategy. This seems especially promising when the global error grows exponentially. In this case the local errors of previous steps have an exponential effect in the global error, so one can think that it makes no sense maintaining the local error below a fixed prescribed tolerance, because the increase of the global error in the new step is mainly due to the propagation of previous errors. It then seems that a gain in efficiency could be obtained by increasing the tolerance of the local error as the propagation of previous errors grows.

From (1.3), and assuming  $e_{n-1}$  is sufficiently small, the error in the  $n$ th step satisfies that

$$(6.1) \quad e_n \simeq R_{n,n-1}e_{n-1} + \delta_n,$$

where  $\delta_n = \delta(y_{n-1}, h_n) = \psi_{h_n}(y_{n-1}) - \phi_{h_n}(y_{n-1})$ , and

$$R_{n,j} = \frac{\partial \phi_{t_n - t_j}}{\partial y}(y(t_j)).$$

In order to simplify the exposition, we hereafter will assume that the equality  $e_n = R_{n,n-1}e_{n-1} + \delta_n$  holds exactly. Due to the properties of the flow, it holds that  $R_{n,j} = R_{n,k}R_{k,j}$  for  $n \geq k \geq j$ , which implies that

$$(6.2) \quad e_n = R_{n,j}e_j + \sum_{k=j+1}^n R_{n,k}\delta_k, \quad 1 \leq j < n.$$

Our aim is to find a condition for the local error  $\delta_n$  that guarantees that the effect in the global error is comparable to the effect of the propagation of previous local errors.

LEMMA 6.1. *Under the assumption that (6.2) holds, if*

$$(6.3) \quad \|\delta_k\| \leq \frac{h_k \|R_{k,j}\| \varepsilon_j}{C}, \quad \varepsilon_j = \frac{\|e_j\|}{t_j - t_0}, \quad C = \max_{j < k < n} \left( \frac{\|R_{n,k}\| \|R_{k,j}\|}{\|R_{n,j}\|} \right),$$

then the following holds for  $1 \leq j < n$

$$\varepsilon_n \leq \|R_{n,j}\| \varepsilon_j.$$

PROOF. From the hypothesis of the lemma and (6.2) we arrive at the inequality

$$\|e_n\| \leq \|R_{n,j}\| (t_j - t_0) \varepsilon_j + \sum_{k=j+1}^n \|R_{n,k}\| \frac{\|R_{k,j}\| h_k \varepsilon_j}{C},$$

which can be rewritten as

$$\|e_n\| \leq \|R_{n,j}\| \varepsilon_j \left( (t_j - t_0) + \sum_{k=j+1}^n h_k \frac{1}{C} \frac{\|R_{n,k}\| \|R_{k,j}\|}{\|R_{n,j}\|} \right)$$

and by definition of  $C$  we arrive at

$$\|e_n\| \leq \|R_{n,j}\| \varepsilon_j \left( (t_j - t_0) + \sum_{k=j+1}^n h_k \right) = \|R_{n,j}\| \varepsilon_j (t_n - t_0),$$

which leads to the required result.

LEMMA 6.2. *Under the assumptions of Lemma 6.1, if  $\varepsilon_{n-1} \leq \|R_{n-1,j}\| \varepsilon_j$  and  $\|\delta_n\| \leq \frac{1}{C^2} \|R_{n,n-1}\| h_n \varepsilon_{n-1}$ , then it holds that*

$$(6.4) \quad \varepsilon_n \leq \|R_{n,j}\| \varepsilon_j.$$

PROOF. Using both inequalities in the hypothesis of the lemma, we arrive at

$$\|\delta_n\| \leq \frac{1}{C^2} \|R_{n,n-1}\| h_n \|R_{n-1,j}\| \varepsilon_j = \frac{1}{C} \frac{\|R_{n,n-1}\| \|R_{n-1,j}\|}{C} h_n \varepsilon_j,$$

and taking into account that, by definition of  $C$ ,

$$\frac{\|R_{n,n-1}\| \|R_{n-1,j}\|}{C} \leq \|R_{n,j}\|$$

holds, the first inequality in (6.3) is obtained, which leads to the required inequality.

REMARK 6.1. The parameter  $C$  depends on the problem to be solved, and it always hold that  $C \geq 1$ . In a linear problem of the form  $y' = Ay$ , it holds that  $C = 1$  if  $A$  is a symmetric matrix. However,  $C$  can be considerably higher, even for linear problems.

The standard way to adjust the step size depends on a fixed tolerance  $\tau$  provided by the user and on a norm  $\|\cdot\|$  in which the local error is measured. The standard approach is based on the assumption that the ideal step-size at the  $n$ th step is the largest  $h_n$  such that the norm of the local error satisfies

$$(6.5) \quad \|\delta_n\| \leq h_n \tau.$$

See [7] (pp. 334–340) and [5] (p. 167) for technical details on the variable step-size strategies actually used in practice. By considering (6.2) with  $j = 0$ , condition (6.5) leads the bound

$$(6.6) \quad \|e_n\| \leq \tau \sum_{k=1}^n \|R_{n,k}\| h_k.$$

It is worth noting that condition (6.5) does not guarantee that the global error at a certain time  $t$  is below the prescribed tolerance. One rather expect that in general the global error will decrease proportionally with the tolerance  $\tau$  (the bound (6.6) may give a hint of this). In practice, the norm of the exact local error is not available, and (often very crude) estimations of the local error are used instead.

We propose an alternative variable step-size strategy based on Lemma 2 as follows. Choose  $h_1$  as in the standard way, by requiring that (6.5) holds, and for  $n \geq 2$ , determine  $h_n$  so that

$$(6.7) \quad \|\delta_n\| \leq \frac{h_n \varepsilon_{n-1}}{C^2},$$

holds, where  $\varepsilon_{n-1}$  and  $C$  are given in (6.3). As  $\|R_{n,n-1}\| \leq 1$ , Lemma 6.2 can be applied, which in particular gives  $\varepsilon_n \leq \|R_{n,1}\| \varepsilon_1$ . Since  $h_1$  has been chosen so that  $\|\varepsilon_1\| \leq \tau$ , we get the bound

$$(6.8) \quad \|e_n\| \leq \tau \|R_{n,1}\| (t_n - t_0).$$

Compared to the bound (6.6) of the global error obtained for the standard variable step-size strategy, the global error admits, when the new strategy is applied, a reasonable bound. Note that the ratio between these two bounds grows at most linearly with time, while the new strategy will allow applying larger step-sizes as the integration proceeds (provided that the sequence  $\varepsilon_n = e_n/(t_n - t_0)$  is increasing).

Condition (6.7) can then be interpreted as (6.5) with a variable tolerance  $\tau = \varepsilon_{n-1}/C^2$ . We then propose modifying the standard variable step-size strategy by replacing the fixed tolerance  $\tau$  in (6.5) by  $\max(\tau, K\varepsilon_{n-1})$ , where  $K$  is a fixed parameter aimed to replace  $1/C^2$ , where  $C$  is given by (6.3). As  $C \geq 1$ ,  $K$  must be chosen so that  $0 < K \leq 1$ .

If we take a value of  $K = 0$ , the new strategy will become the standard strategy because in that case  $\max(\tau, K\varepsilon_{n-1}) = \tau$ . If  $K$  is too large, so that  $K > 1/C^2$  and (6.7) does not hold, then there is no guarantee that the bound (6.8) holds, which may result in an unacceptably fast growth of the global error  $\|e_n\|$ .

Unfortunately, the optimal parameter  $K \in (0, 1]$  depends on the problem to be solved. In order to implement our new variable step-size strategy in a robust way, it would be desirable to devise some strategy to adjust the value of  $K$  automatically as the integration of the problem proceeds. Such procedure should in principle choose small values of  $K$  (or even  $K = 0$ ) for problems with nonincreasing or slowly increasing global error, and values close to  $K = 1$  for problems exhibiting exponential growth of the global error. This is beyond the scope of the present paper, and might be the subject of a future work.

## 7 Numerical experiments for the new variable step-size strategy.

We have implemented the new strategy based on (6.7) in our experimental code. We have performed several numerical experiments for different initial value problems, and we present the results obtained for four of them: The first two problems in Section 5, the initial value problem referred as *Lorenz*, taken from [5] (pp. 120–121) and the problem referred as *Twobody* taken from [7] (p. 121) with eccentricity = 0.5.

In order to test the efficiency of the integrations, we have shown plots of the number of steps needed in the integration versus the global error at the end of the interval of integration, in double logarithmic scale. Each efficiency diagram is obtained by numerically solving the problem with different local error tolerances. Several diagrams are shown in each figure: One corresponds to the application of our method with the usual step-size strategy ( $K = 0$ ), and the rest of the curves correspond to the new strategy, each one with a different value of the parameter  $K$  ( $K = 0.1, 0.5, 1$ ).

We can see the results obtained for the *Arenstorf* problem in Figure 7.1. The continuous line corresponds to the usual step-size strategy, and in all cases the new strategy gets better results. The optimal value of  $K$  among the three we have tried is  $K = 0.5$ . The gain in efficiency for this problem in the case of  $K = 0.5$  is around 33% for all tolerances. Similar efficiency improvement is

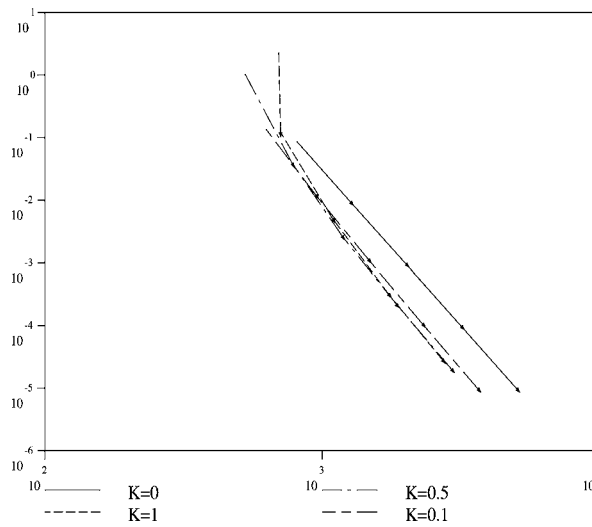


Figure 7.1: The comparison of the cost for the *Arenstorf* problem between the new step-size strategy and the usual strategy.

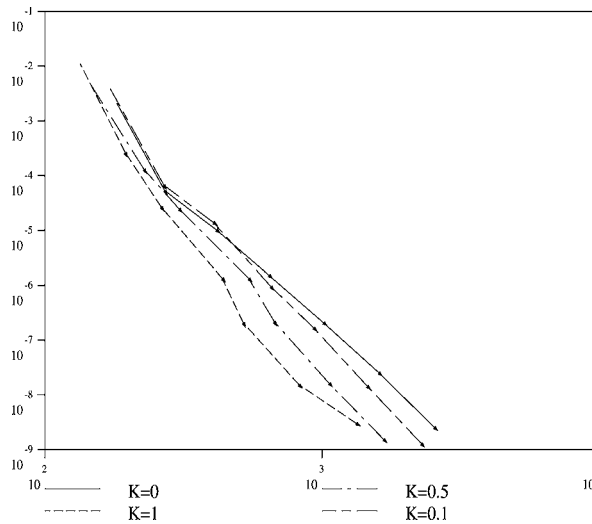
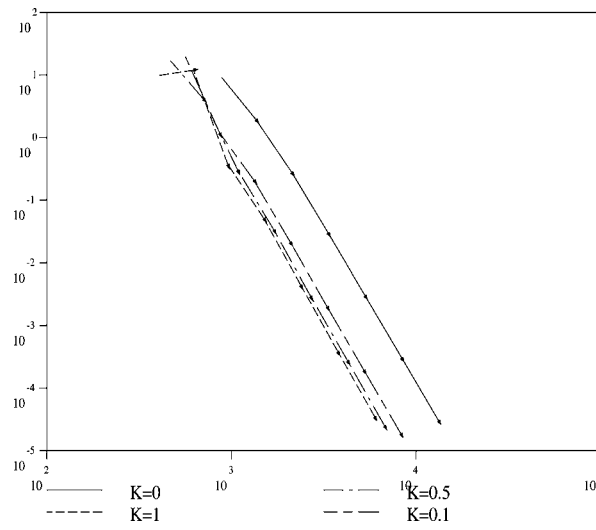
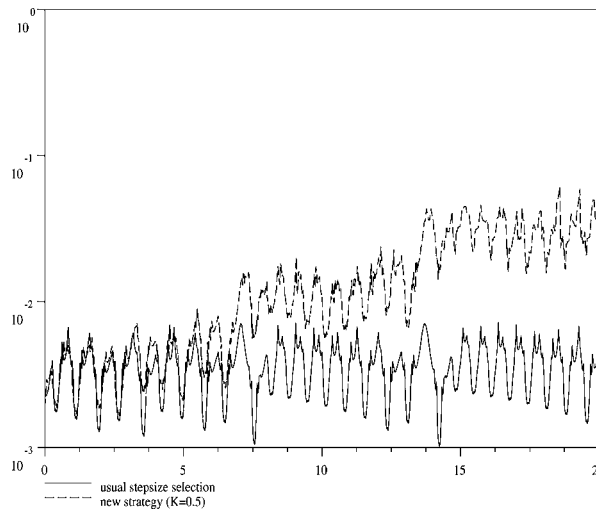


Figure 7.2: Efficiency for the *Pleiades* problem with usual step-size strategy and the new step-size strategy.

obtained for  $K = 0.1$  and  $K = 1$  with the exception of the integration with the largest tolerance with  $K = 1$ .

Figure 7.2 shows the difference between standard step-size strategy and the new step-size strategy for the *Pleiades* problem. The best efficiency results are obtained for  $K = 1$ . Again the continuous line corresponds to the usual strategy and the rest have been obtained using the new strategy. The gain in efficiency for  $K = 1$  is between 20% and 45%, depending on the tolerance.

Figure 7.3: Efficiency comparison for the *Lorenz* problem.Figure 7.4: Step-size versus time in the integration of the *Lorenz* problem.

The results for the *Lorenz* problem are shown in Figure 7.3. It can be seen that applying the new strategy we need fewer steps to get the same accuracy: for  $K = 1$  or  $K = 0.5$  only 55% of steps are needed, so the gain of efficiency is 45%. In fact, we have observed that, in this problem, despite the substantial reduction of the number of steps obtained with our variable step-size strategy, the global error essentially does not change. In this example, incrementing the value of the parameter  $K$  the integration needs fewer steps without compromising the precision of the results. In Figure 7.4, we compare the step-sizes used when

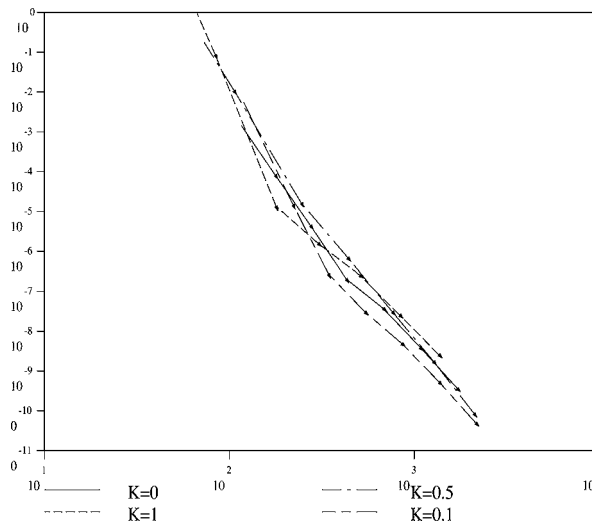


Figure 7.5: Efficiency comparison for the *Twobody* problem.

integrating the Lorenz problem with the standard variable step-size strategy ( $K = 0$ ) and the new strategy (with  $K = 0.5$ ).

Note that in the three problems presented above the error grows very fast, and so the new strategy gives very good results. The new strategy does not seem to improve efficiency for problems whose error does not grow fast. Figure 7.5 shows the efficiency diagram obtained with the *Twobody* problem with eccentricity equal to 0.5. In this case, the new strategy acts like a change of tolerance, there is no gain of efficiency.

We have seen that the new step-size strategy can be used to get a gain in efficiency but one can wonder whether the estimation of the global error may be affected by the choice of a different step-size sequence. We have performed extensive numerical experiments to test that issue, and we have observed that the quality of the estimation of the global error is not affected by the new step-size strategy (the figures thus obtained are very similar to those in Section 5). It should be noted that the theoretical motivation given in Sections 2 and 3 for the proposed global error estimation procedure does not depend on the actual step-size sequence  $\{h_n\}$  used to integrate numerically the problem, but rather it is only required that each  $h_n$  be sufficiently small so that the estimate (3.1) makes sense.

## REFERENCES

1. J. R. Bunch, *The weak and strong stability of algorithms in numerical linear algebra*, Linear Algebra Appl., 88/89 (1987), pp. 49–66.
2. J. C. Butcher, *The Numerical Analysis of Ordinary Differential Equations*, Willey, New York, 1987.
3. J. R. Dormand, J. P. Gilmore, and P. J. Prince, *Globally embedded Runge–Kutta schemes*, Ann. Numer. Math., 1 (1994), pp. 97–106.

4. J. R. Dormand and P. J. Prince, *A family of embedded Runge–Kutta formulae*, J. Comp. Appl. Math., 6 (1980), pp. 19–26.
5. E. Hairer, S. P. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations I. Non-Stiff Problems*, Springer-Verlag, 2nd edn, 1993.
6. A. Murua and J. Makazaga, *Cheap one-step global error estimation for ODEs*, New Zealand J. Math., 29 (2000), pp. 211–221.
7. L. F. Shampine, *Numerical Solution of Ordinary Differential Equations*, Chapman and Hall, 1994.
8. W. M. Lioen and J. J. B. de Swart, *Test set for initial value problem solvers*, Release 2.1, September 1999, <http://www.cwi.nl/cwi/projects/IVPtestset/>.