Continuous Numerical Methods for ODEs with Defect Control ⋆

W.H. Enright

University of Toronto, Toronto, Ontario, Canada M5S 3G4

Abstract

Over the last decade several general purpose numerical methods for ordinary differential equations (ODEs) have been developed which generate a continuous piecewise polynomial approximation that is defined for all values of the independent variable in the range of interest. For such methods it is possible to introduce measures of the quality of the approximate solution based on how well the piecewise polynomial satisfies the ODE. This leads naturally to the notion of 'defect control'. Numerical methods that adopt error estimation and stepsize selection strategies in order to control the magnitude of the associated defect can be very effective and such methods are now being widely used.

In this paper we will review the advantages of this class of numerical methods and present examples of how they can be effectively applied. We will focus on numerical methods for initial value problems (IVPs) and boundary value problems (BVPs) where most of the developments have been introduced but we will also discuss the implications and related developments for other classes of ODEs such as delay differential equations (DDEs) and differential algebraic equations (DAEs).

1 Introduction

Consider the ordinary differential equation

$$y' = f(t, y), \tag{1}$$

with exact solution denoted by $y(t)$. Traditional discrete numerical methods partition the interval of interest, $[t_0, t_F]$, by introducing a mesh $t_0 < t_1, \cdots < t_N = t_F$ and generate a discrete approximation $y_i \approx y(t_i)$ for each associated

⋆ This research was supported by the Natural Science and Engineering Research Council of Canada and by Communication and Information Technology Ontario.

meshpoint. The number of meshpoints, N, and the distribution of these meshpoints is usually determined adaptively by the method in an attempt to deliver acceptable accuracy at a minimum cost. These methods generally accomplish this objective by keeping N as small as possible subject to a constraint that an indirect measure of $\max_{i=1\cdots N} \|y(t_i) - y_i\|$ be kept small (relative to an accuracy parameter TOL). (We use $\|\cdot\|$ to represent the max-norm for vectors and the induced matrix norm for matrices.) Different methods implement very different strategies in an attempt to achieve this indirect error control and this can make it particularly challenging to interpret the accuracy of a numerical solution.

In recent years the notion of a continuous extension (of an underlying discrete method) has received considerable attention. With this approach, one associates with each discrete approximation, $\{t_i, y_i\}_{i=1}^N$, a piecewise polynomial, $u(t)$, defined for all t in the interval $[t_0, t_F]$ and satisfying $u(t_i) = y_i$ for $i = 1, 2 \cdots N$. (For a detailed discussion of the advantages and costs of generating such extensions see [8], [14], [18], [24] and [27].) For example, consider applying a method based on a standard s-stage, p^{th}-order Runge-Kutta formula to (1). The corresponding discrete approximations will then satisfy

$$y_i = y_{i-1} + h_i \sum_{j=1}^{s} \omega_j k_j \ , \tag{2}$$

where

$$k_j = f(t_{i-1} + h_i c_j, y_{i-1} + h_i \sum_{r=1}^{s} a_{jr} k_r),$$

for $j = 1, 2 \cdots s$; $i = 1, 2 \cdots N$ and $h_i = t_i - t_{i-1}$. A continuous extension is derived by introducing the additional stages, $k_{s+1}, k_{s+2} \cdots k_{\bar{s}}$ and polynomials $b_j(\tau)$ of degree $\leq p$ for $j = 1, 2 \cdots \bar{s}$ such that the polynomial $u_i(t)$ defined by

$$u_i(t) = y_{i-1} + h_i \sum_{j=1}^{\bar{s}} b_j(\frac{t - t_{i-1}}{h_i}) k_j \ , \tag{3}$$

satisfies $u_i(t) = y(t) + O(h^p)$ for $t \in (t_{i-1}, t_i)$ and $h = \max_{i=1}^N h_i$. Formulas of this type are called Continuous Runge-Kutta (CRK) formulas. The polynomials $\{u_i(t)\}_{i=1}^N$ then define a piecewise polynomial, $u(t)$, which will be continuous on $[t_0, t_F]$ and interpolate the underlying discrete approximation if $b_j(1) = w_j$ for $j = 1, 2 \cdots s$ and $b_{s+1}(1) = b_{s+2}(1) \cdots = b_{\bar{s}}(1) = 0$.

In deriving CRK formulas of order p several issues arise which can have a significant impact on the implementation and hence on the performance of the resulting continuous method. If the extra stages are restricted to be 'explicit', that is if

$$k_{s+j} = f(t_{i-1} + c_{s+j}h_i, y_{i-1} + h_i \sum_{r=1}^{s+j-1} a_{s+j,r}k_r)$$

for $j = 1, 2 \cdots (\bar{s} - s)$, then implementation is straightforward and the cost of obtaining the continuous approximation is only the additional $\bar{s} - s$ evaluations of the differential equation on each step. It is therefore generally preferable to derive and implement explicit CRK methods although there are classes of problems where one can only use implicit CRK formulas. Differential algebraic problems, DAEs, are one such class and we will consider them in more detail in section 5.

Another issue that can be important is the magnitude of the defect or residual that is associated with the continuous approximation. This quantity can be interpreted as a measure of the quality of the numerical solution. For a piecewise polynomial approximation $u(t)$ associated with the ODE (1) one defines the defect, $\delta(t)$, for $t \in (t_0, t_F)$ by

$$\delta(t) = u'(t) - f(t, u(t)).$$

Note that, in deriving CRK formulas, our assumption that $u(t)$ be order p implies, for sufficiently smooth problems,

$$u'(t) = y'(t) + O(h^{p-1}). \tag{4}$$

Furthermore, since

$$\begin{aligned} \delta(t) &= u'(t) - f(t, u(t)) - y'(t) + f(t, y(t)) \\ &= (u'(t) - y'(t)) + (f(t, y(t)) - f(t, u(t))), \end{aligned} \tag{5}$$

then for differential equations that are Lipschitz continuous, $\|\delta(t)\|$ will be at worst $O(h^{p-1})$.

If we let $z_i(t)$ be the solution of the local initial value problem

$$z_i' = f(t, z_i), \quad z_i(t_{i-1}) = y_{i-1},$$

for $i = 1, 2 \cdots N$; then the local error, $le_i(t)$, associated with the continuous approximation on step i is defined for $t \in (t_{i-1}, t_i)$ to be

$$le_i(t) = z_i(t) - u_i(t).$$

It is well known that the discrete local error of a p^{th}-order Runge-Kutta formula (2) must be $O(h_i^{p+1})$. That is

$$\begin{aligned} le_i(t_i) &= z_i(t_i) - u_i(t_i) \\ &= z_i(t_i) - y_i \\ &= O(h_i^{p+1}). \end{aligned} \tag{6}$$

If we derive order p CRK formulas which satisfy the additional constraint that the associated local error be $O(h_i^{p+1})$ for $t \in (t_{i-1}, t_i)$ then we will have

$$u_i'(t) = z_i'(t) + O(h_i^p) \tag{7}$$

and therefore, for $t \in (t_{i-1}, t_i)$,

$$\begin{aligned} \delta(t) &= (u_i'(t) - z_i'(t)) + (f(t, z_i(t)) - f(t, u_i(t))) \\ &= O(h_i^p). \end{aligned} \tag{8}$$

Furthermore, if we derive order p CRK formulas with the stronger additional constraint that, for $t \in (t_{i-1}, t_i)$, the local error be $O(h_i^{p+1})$ and satisfy

$$le_i(t) = \psi_i(\tau)\; D(t_{i-1})\; h_i^{p+1} + O(h_i^{p+2}), \tag{9}$$

where $\tau = \frac{t - t_{i-1}}{h_i}$, $D(t)$ is a function depending only on the problem, and $\psi_i(\tau)$ is a polynomial in τ whose coefficients are independent of the problem and the stepsize, it can be shown from (8) (see [3] for details) that

$$\delta(t) = \psi_i'(\tau)\; D(t_{i-1})\; h_i^p + O(h_i^{p+1}). \tag{10}$$

In this paper we are considering continuous methods which are designed to directly monitor and control the maximum magnitude of an estimate of the defect of the piecewise polynomial $u(t)$ that is delivered as the approximate solution. We will focus on methods based on an order p CRK formula but most of the discussion and analysis will apply to other classes of continuous methods such as those based on multistep formulas. Note that an order p CRK will always satisfy (5) but, without additional constraints, such formulas may be more difficult to implement in an effective and reliable way (when the objective is to control the magnitude of $\|\delta(t)\|$) since:

- The magnitude of the associated defect will generally only be $O(h_i^{p-1})$,
- Although the defect can be sampled at any point in the interval of interest it may not be easy to justify a rigorous, inexpensive estimate of its maximum magnitude over each step.

The first of these difficulties can be overcome by considering only order p CRK formulas with local error that is $O(h_i^{p+1})$ and both difficulties can be overcome by considering only those order p CRK formulas that satisfy (9). In this latter case as $h_i \to 0$, for any $t \in (t_{i-1}, t_i)$, $\delta(t)$ will satisfy (10) and therefore for $\bar{t} \in (t_{i-1}, t_i)$ (corresponding to a local maximum of $|\psi_i'(\frac{t-t_{i-1}}{h_i})|$), $\|\delta(\bar{t})\|$ will be an asymptotically correct estimate of the maximum magnitude of the defect on the i^{th} step. Note that since $\psi_i'(\tau)$ is a polynomial which is independent of the problem and the stepsize, the location of its local maximum magnitude, $\bar{\tau}$, (for $\tau \in (0,1)$) is known and the corresponding value for $\bar{t}$ is $\bar{t} = t_{i-1} + \bar{\tau} h_i$.

In the next section we will consider continuous methods for IVPs based on defect control and in subsequent sections we will consider such methods for BVPs and DDEs. In each of these areas there are some general purpose software packages available. Finally we will consider the development of methods for DAEs based on defect control. We will discuss some prototype and experimental codes that implement this approach.

2 Initial Value Methods

The development of software based on defect control for the numerical solution of IVPs in ODEs has a history that goes back several decades and is closely related to the notion of backward error analysis. Consider the standard IVP

$$y' = f(t, y), \quad y(t_0) = y_0, \quad t \in [t_0, t_F]. \tag{11}$$

Hull [19] and Stetter [25] investigated the reliability (or 'effectiveness') of various error control strategies for discrete methods applied to (11) by establishing conditions which would guarantee the existence of a piecewise approximating function, $\hat{u}(t) \in C^0[t_0, t_F]$, which interpolates the discrete approximate solution y_i and satisfies a slightly perturbed IVP:

$$\hat{u}' = f(t, \hat{u}) + \hat{\Delta}(t), \ \hat{u}(t_0) = y_0, \tag{12}$$

where $\hat{\Delta}(t) \in C^0[t_0, t_F]$ and satisfies

$$\|\hat{\Delta}(t)\| \le \hat{k} \ \ \mathrm{TOL}\,, \tag{13}$$

for some modest value of $\hat{k}$ (independent of both the problem and the method), and TOL is the specified error tolerance. In these investigation $\hat{u}(t)$ and $\hat{\Delta}(t)$ were generally not computable but one could use standard results from mathematics, such as the Grobner-Alekseev variation of constants formula (see [23] for details), to obtain appropriate global error bounds. For example if (11), (12) and (13) are satisfied then it is straightforward to show

$$y(t_i) - y_i = y(t_i) - \hat{u}(t_i) = \int_{t_0}^{t_i} K(t,s)\hat{\Delta}(s)ds, \qquad (14)$$

where $K(t,s)$ is a variational matrix that depends only on the problem, and this implies

$$\|y(t_i) - y_i\| \leq (t_i - t_0)\hat{k}\ K_{max}\ \mathrm{TOL}\ , \qquad (15)$$

where K_{max} is a bound on $\|K(s,t)\|$. Note that this is an example of backward error analysis where the computed solution is guaranteed to exactly satisfy a slightly perturbed IVP and one can interpret K_{max} as a type of condition number for the problem (which quantifies how sensitive the solution can be to small changes in the problem specification).

Subsequently several investigators, who were primarily interested in dense output (or off-mesh approximations), analysed and developed computable piecewise polynomial interpolants, $u(t)$, which could be efficiently implemented with new or existing discrete methods (see for example [26], [18], [24], [28] and [8]). It was soon recognized that, once $u(t)$ was generated, the corresponding defect could be sampled and new reliable error and stepsize control strategies could be developed with the objective of directly satisfying relationships analogous to (12), (13).

As we have noted earlier, when one considers CRK formulas, the requirement that $\|\delta(t)\|$ be of optimal order and easy to bound by an inexpensive error estimate (for $t \in (t_{i-1}, t_i)$) will generally impose additional constraints on what would comprise a suitable local interpolating polynomial, $u_i(t)$. In a series of investigations ([1], [2], [3], [15], [16] and [17]), several order p CRK formulas have been developed and compared. These investigations have also considered the relative advantages of several alternative defect control strategies for $3 \leq p \leq 8$. We will now consider two of the most promising strategies.

The first strategy assumes that the local interpolant $u_i(t)$ defined by (3) satisfies (7) and that the estimate of the maximum magnitude of the defect is obtained by the heuristic

$$\begin{aligned} est_i &= \|\delta(t_{i-1} + \hat{\tau}h_i)\| \\ &= \|u_i'(t_{i-1} + \hat{\tau}h_i) - f(t_{i-1} + \hat{\tau}h_i, u_i(t_{i-1} + \hat{\tau}h_i))\| \end{aligned} \qquad (16)$$

where $\hat{\tau}$ is chosen in a careful way (see [1] for a detailed discussion of how $\hat{\tau}$ can be chosen). We will refer to this strategy as the 'relaxed' defect control

Formula	p	s	$\bar{s}$	$\tilde{s}$
CRK4	4	4	6	7
CRK5	5	6	9	11
CVSS6B	6	7	11	14
CVSS7	7	9	15	20
CVSS8	8	13	21	28

Table 1
Cost per step of relaxed and strict defect control for some CRK formulas

strategy. This heuristic for controlling the maximum magnitude of the defect works well on most problems but the strategy is not asymptotically justified (as $h \rightarrow 0$) and it can severely underestimate the size of the defect for some problems. Table 1 reports for $4 \leq p \leq 8$ the value of s and $\bar{s}$ for some order p CRK which have been found to be particularly effective. The higher order formulas are members of families of formulas derived in [27] (the particular coefficients are identified in [3]). Note that the cost per step of a method using this error control strategy for these particular CRKs is $\bar{s}$ derivative evaluations. This follows since although one must perform an additional derivative evaluation to sample the defect (at $t_{i-1}+\hat{\tau}h_i$), each of the formulas is designed to ensure that $u(t) \in C^1[t_0, t_F]$ by requiring that $u_i'(t_{i-1}+h_i) = f(t_{i-1}+h_i, y_i)$. This implies that, on all but the first step, k_1 will be available (as an internal stage of the previous step).

A more rigorous error control strategy (which we will refer to as the 'strict' defect control strategy) for the same set of underlying discrete formulas can be developed by requiring that the corresponding continuous extension satisfy (9) as well as (7). With this additional constraint we have available (as discussed earlier) an asymptotically correct estimate of the maximum magnitude of the defect (on the i^{th} step) given by $est_i = \|\delta(t_{i-1} + \bar{\tau}h_i)\|$ where $\bar{\tau}$ is fixed and independent of the problem or stepsize. One way to generate such a CRK (although it may not be optimal) is to begin with an $\bar{s}$-stage order p CRK, $u_i(t)$, satisfying (3) and (7) and replace the 'extra' stages, $k_{s+j},\ j = 1 \cdots (\bar{s}-s)$ (whose corresponding $c_{s+j} \neq 1$) with the more accurate values

$$\tilde{k}_{s+j} = \begin{cases} f(t_{i-1} + c_{s+j}h_i, u_i(t_{i-1} + c_{s+j}h_i)) & \text{if } c_{s+j} \neq 1, \\ k_{s+j} & \text{if } c_{s+j} = 1. \end{cases} \tag{17}$$

A new, more suitable interpolant, $\tilde{u}_i(t)$ can then be defined by

$$\tilde{u}_i(t) = y_{i-1} + h_i \sum_{j=1}^{s} b_j(\frac{t - t_{i-1}}{h_i})k_j + \sum_{j=1}^{\bar{s}-s} b_{j+s}(\frac{t - t_{i-1}}{h_i})\tilde{k}_{s+j}. \tag{18}$$

It can easily be shown that $\tilde{u}_i(t)$ will satisfy (7), (9) and, when $c_j = 1$ for one value of j in the range $1 \leq j \leq (\bar{s} - s)$ (as is the case for each of the CRK formulas identified in Table 1), it will be an $\tilde{s}$-stage order p CRK with $\tilde{s} = s + 2(\bar{s} - s) - 1$. Table 1 also reports the value of $\tilde{s}$ for this strict defect control strategy using the CRK corresponding to (18).

Clearly a trade-off between efficiency and reliability needs to be addressed when choosing which defect control strategy should be used to solve a particular problem. Fortunately it is convenient and straightforward to implement a numerical method which can apply either strategy (using either (3) with $\hat{\tau}$ or (18) with $\bar{\tau}$ to define the respective interpolants and defect estimates) and thus the choice can be an option which can be selected by the user. From Table 1 we can observe that the cost per step of using the strict defect control strategy can be between 20% and 33% more than that for the relaxed strategy but better control of the size of the defect should be realised.

To illustrate and quantify this trade-off we have run both versions of a numerical method based on the CRK formula CVSS6B, on the 25 standard non-stiff problems of the DETEST package ([12]) at nine error tolerances and assessed how well the defect was controlled. The performance on a typical problem, problem B4, is summarised in Table 2. The statistics reported in this table are:

TOL The specified error tolerance.
Time The computer time required to solve the problem measured in seconds on a SUN Sparc4.
FCN The number of derivative evaluations required to solve the problem.
Steps The number of time steps required to solve the problem.
GL Err The maximum observed global error measured in units of TOL and determined by measuring the global error at 100 equally spaced points per step.
Max Def The maximum magnitude of the defect measured in units of TOLand determined by sampling the defect at 100 equally spaced points per step.
% Succ The percentage of steps where the magnitude of the defect is less than TOL .

These results are typical of that observed on all problems. Both methods are robust in that they are able to deliver, over a wide range of tolerances, a close and consistent relationship between the size of the defect and the specified error tolerance.

Both versions of CVSS6B required over 13,500 steps to solve all of the problems at all tolerances. With the strict defect control strategy the maximum magnitude of the defect rarely (on fewer than 1.8% of the steps) exceeded TOL and never exceeded 7 TOL . With the relaxed defect control strategy

Strategy	TOL	Time	FCN	Steps	GL Err	Max Def	% Succ
Relaxed defect control	10^{-2}	.010	166	15	4.7	1.9	73
	10^{-3}	.015	254	22	18.0	1.2	91
	10^{-4}	.023	375	32	28.4	1.3	84
	10^{-5}	.031	507	46	34.5	1.7	74
	10^{-6}	.046	749	68	37.6	1.5	62
	10^{-7}	.067	1101	100	39.0	1.5	49
	10^{-8}	.099	1618	147	40.5	1.5	45
	10^{-9}	.145	2377	216	41.1	1.5	45
Strict defect control	10^{-2}	.014	253	17	1.2	.88	100
	10^{-3}	.021	379	24	9.5	.94	100
	10^{-4}	.027	491	35	15.6	.98	100
	10^{-5}	.041	743	53	16.6	.76	100
	10^{-6}	.061	1093	78	16.6	.66	100
	10^{-7}	.090	1625	116	16.5	.61	100
	10^{-8}	.131	2381	170	16.5	.58	100
	10^{-9}	.193	3501	250	16.5	.56	100

Table 2
Performance of CVSS6B on problem B4 of DETEST

the maximum magnitude of the defect exceeded TOL on 20% of the steps but it rarely exceeded 5 TOL (on fewer than 1.2% of the steps) and it never exceeded 18 TOL .

An alternative rigorous defect control strategy based on the use of a different norm has been proposed and justified in [20]. On each step one introduces a weighted L_2-norm (which can be interpreted as an average magnitude of the defect),

$$\|\delta_i(t)\|_2 = 1/h_i \left(\int_{t_{i-1}}^{t_i} \|\delta_i(s)\|^2 \, ds \right)^{1/2} , \tag{19}$$

A method can then be developed with an error control strategy that attempts to ensure that

$$\|\delta_i(t)\|_2 \leq \mathrm{TOL} \, . \tag{20}$$

As is pointed out in [20], $\delta_i(t)$ is usually known at the meshpoints and therefore, for sufficiently smooth problems, one can derive a low cost, asymptotically correct estimate of $\|\delta_i(t)\|_2^2$ using a suitably chosen Lobatto quadrature formula (which would require only a few additional evaluations of the defect). This approach has been implemented and shown to be very effective for a large class of problems.

3 Boundary Value Methods

Numerical methods for BVPs of the form

$$y' = f(t, y), \ \ t \in [t_0, t_F], \ \ g(y(t_0), y(t_F)) = 0. \tag{21}$$

generally produce a discrete approximation on a mesh $t_0 < t_1 < \cdots < t_N = t_F$ by solving a large coupled nonlinear system of equations. If the underlying formula that determines the discrete solution is a Runge-Kutta or collocation formula then it is straightforward to introduce a continuous extension $u(t)$ and the associated defect $\delta(t)$ (as we have done for IV methods). From Table 1 we see that the cost per step to compute $u(t)$ and estimate the size of the corresponding defect can be as great as applying the underlying discrete formula. For BV methods the cost per step to determine $u(t)$ and $\delta(t)$, after the discrete solution has been computed, remains the same while the cost of solving for the discrete solution is generally much greater. A consequence of this is that once a converged discrete solution is determined by a BV method (based on the use of a CRK or collocation formula), a continuous extension with an optimum $O(h^p)$ defect can be computed at very little incremental cost.

When these formulas are used to determine the discrete solution, defect-based error control and mesh-refinement strategies can be particularly attractive. This approach has been followed in the development of the methods MIRKDC [10] and bvp4c [20] which have been found to be effective for solving a wide class of problems.

In the numerical solution of BVPs one often encounters difficulties with convergence of the iteration scheme that is used to solve the nonlinear system associated with the discrete mesh. This can be the result of a poor choice of mesh and/or a poor initial guess for the discrete solution. In either case, if the method has available a piecewise polynomial approximation $\bar{u}(t)$ with an associated defect $\bar{\delta}(t)$ (as would be the case, for example, if $\bar{u}(t)$ were associated with a mesh and previously computed discrete solution that was judged not to be sufficiently accurate), then these deficiencies can often be overcome by using the size of the defect to help guide the mesh refinement and using $\bar{u}(t)$ to generate the required initial guess. With this approach one can also use the estimates of the maximum magnitude of the defect to help ensure that the approximate solution that is ultimately delivered by the method, $u(t)$, satisfies

$$\|\delta(t)\| = \|u'(t) - f(t, u(t))\| \leq \text{ TOL }, \tag{22}$$

and

$$g((u(t_0), u(t_F)) = 0.$$

Note that, with this approach, one could consider using inexpensive interpolants for the mesh refinement strategies, and the more expensive rigorous interpolants for assessing the accuracy of the numerical solution.

When such strategies are adopted by a method, one only has to compute the interpolant and defect estimate on the final iteration after the underlying discrete approximation has converged. Intermediate calculations associated with preliminary coarse meshes or initial iterations (before convergence) either would not require the determination of any interpolant or would only require the less expensive relaxed interpolant.

Numerical experience reported in [10] and [20] shows that BV methods that implement such defect based strategies can outperform methods based on more traditional strategies especially when a strongly non-uniform mesh is appropriate. Even on problems where asymptotic analysis is not necessarily relevant, carefully designed defect-based strategies can quickly lead to a suitable mesh and rapid convergence on that mesh.

4 Delay Differential Equation Methods

A class of numerical methods based on CRK formulas with defect control has been analysed [6] for systems of retarded and neutral DDEs of the form,

$$y' = f(t, y, y(t - \sigma_1(t, y(t))), y'(t - \sigma_2(t, y(t)))), \ \ t \in [t_0, t_F]. \tag{23}$$

$$y(t) = \phi(t), \ \ t \leq t_0, \tag{24}$$

where $\sigma_1(t, y)$ and $\sigma_2(t, y)$ are positive scalar functions. One particular sixth order formula from this class (the formula CVSS6B discussed in section 2) has been implemented in a software package, DDVERK [4], for these DDEs.

For this class of problems a discrete method must be able to approximate the solution at off-mesh points in order to evaluate the differential equation at an arbitrary point, $t \in [t_0, t_F]$. Therefore the requirement that the numerical solution be a piecewise polynomial, $u(t)$, does not impose any extra cost and one can associate an approximation $u(t)$ (with corresponding defect $\delta(t)$) with any numerical method.

To be effective for this class of problems a numerical method must be able to detect and efficiently handle the discontinuities that inevitably arise and are propagated as the integration proceeds. Automatic techniques for detecting and accurately crossing points of discontinuity for standard IVPs based on monitoring changes in the associated defect have been proposed and justified in [9]. This technique has been adapted and refined for DDEs in the solver DDVERK (see [4] for a discussion of the details) where it has proved to be very effective for a wide class of problems. It is certainly competitive with the alternative strategy which explicitly checks for discontinuities at all possible locations where propagation is possible. This is particularly true for systems of equations with multiple delays where the number of potential points of discontinuity can be quite large relative to the number of significant or relevant points of discontinuity. The defect based strategy for coping with discontinuities essentially adjusts the stepsize selection (as well as the error control mechanism) only on those steps where the presence of a point of discontinuity has severely reduced the stepsize.

5 Differential Algebraic Equation Methods

In recent years there has been considerable interest and progress made in the development of numerical methods for special classes of DAEs. Nevertheless, very few methods can be applied directly to a system of DAEs in the most general form

$$F(t, y, y') = 0, \;\; y(t_0) = y_0, \;\; t \in [t_0, t_F], \tag{25}$$

with $\frac{\partial F}{\partial y'}$ known to be singular. Note that if this matrix is nonsingular for all $t \in [t_0, t_F]$ the problem is said to have index 0. In this case one can solve the nonlinear system associated with (25) to determine $y'(t)$ for any prescibed value of t and $y(t)$. Any initial value method can be applied and special DAE methods are not necessary.

In general the 'index' of a problem of the form (25) refers to the minimum number of times one has to differentiate the equation, $F(t, y, y') = 0$, in order to derive an equivalent initial value problem where $y'(t)$ can be determined uniquely in terms of $t, y(t), F$ and various partial derivatives of F. The higher the index of a problem, the more sensitive the solution can be to perturbations in the data and the more difficult it becomes to develop reliable numerical methods. Currently there are several reliable general purpose numerical methods for index 1 problems and other reliable methods designed for special classes of index 2 and index 3 problems.

The DAEs that arise in application areas, such as the modelling of constrained mechanical systems or the design of electrical circuits often are of index 2 or index 3 but they possess special structure and numerical methods which exploit this structure have been developed and have received wide acceptance. For example, the algebraic constraints can often be explicitly identified and the system decoupled, $y(t) = [y_1(t), y_2(t)]^T$, and written in the semi-explicit form

$$
\begin{aligned}
y_1'(t) &= f(t, y_1(t), y_2(t)), && (26)\\
0 &= g(t, y_1(t), y_2(t)). && (27)
\end{aligned}
$$

When one considers the development of defect-based error control for DAE methods two key questions must first be answered:

(1) How does one define a sufficiently accurate continuous extension, $u_i(t)$, of the discrete approximation (for $t \in [t_{i-1}, t_i]$)?
(2) What measure of the size of the defect is appropriate to control? That is, can one introduce a measure $\mu_i(\delta)$ such that for $t \in [t_{i-1}, t_i]$ the condition that $\mu_i(\delta(t)) \leq \mathrm{TOL}$ will ensure that the global error will be proportional to TOL and $\mu_i(\delta)$ will be inexpensive to estimate on each step?

These questions were considered in [22] where defect-based error control strategies suitable for important classes of index 2 and index 3, semi-explicit problems were introduced and justified. The approach that was introduced can be applied with any discrete, order p, implicit Runge-Kutta formula to generate, on each step, interpolating polynomials $u_i(t)$ and $v_i(t)$ that approximate $y_1(t)$ and $y_2(t)$ respectively. If one defines the defect of the resulting vector of piecewise polynomials associated with $u(t)$ and $v(t)$ we have (from (26) and (27))

$$
\begin{aligned}
\delta_1(t) &= u'(t) - f(t, u(t), v(t)), && (28)\\
\delta_2(t) &= g(t, u(t), v(t)). && (29)
\end{aligned}
$$

The global errors $\|y_1(t) - u(t)\|$ and $\|y_2(t) - v(t)\|$ were analysed and shown to be bounded by a suitable multiple of TOL provided $\delta_1(t), \delta_2(t)$, and $\delta_2'(t)$ were all suitably bounded in norm. Corresponding measures $\mu_i(\delta)$ were proposed and associated estimates introduced which could be the basis for an effective defect-based numerical method for semi-explicit DAEs.

Another approach has been considered in [21] where no assumptions are made on the structure of the DAE. In order to determine the piecewise polynomial, $u(t)$ which approximates the solution to (25) one begins with an implicit continuous extension of a discrete, order p, implicit Runge-Kutta formula. One then introduces an associated overdetermined system of nonlinear equations on each time step by requiring that the corresponding approximating polynomial, $\tilde{u}_i(t)$ satisfy (in a least squares sense) the defining equations of the underlying continuous extension as well as additional 'collocation' equations (which are equivalent to asking that (25) be satisfied at a prescribed set of sample points). The defect, $\tilde{\delta}(t)$, associated with the resulting piecewise polynomial, $\tilde{u}(t)$, is defined by

$$\tilde{\delta}(t) = F(t, \tilde{u}(t), \tilde{u}'(t)). \tag{30}$$

Conditions on the choice of underlying implicit CRK formulas and on the number and choice of collocation points are identified which result in $\|\tilde{\delta}(t)\|$ being $O(h^p)$ for sufficiently differentiable index 1 and index 2 problems. Estimates of $\|\tilde{\delta}(t)\|$ are justified and an experimental code introduced to illustrate the validity of this approach. A general purpose numerical method based on this approach is under development.

6 Summary and Conclusions

As is clear from our discussion so far there are now several general purpose numerical methods for important classes of ODEs that produce piecewise polynomial approximate solutions and attempt to directly control the magnitude of the associated defect. These methods, although more costly than classical discrete methods, can be efficiently implemented and they produce solutions whose accuracy can be more readily interpreted and compared.

We have also shown that when implementing numerical methods using defect control one must address a trade-off between reliability and efficiency. This trade-off arises from a choice between the use of an inexpensive heuristic or a more expensive (but asymptotically correct) estimate of the maximumum magnitude of the defect. This choice can be left to the user but the implications must be understood when interpreting the numerical results.

There are two difficulties that have not been discussed which limit the applicability of this class of methods and which should be addressed in future investigations. If the underlying problem is not sufficiently smooth, then one is restricted to the use of low order methods and defect control can be less competitive with the more classical approach at low orders. Also, at limiting precision, where the effect of round-off error may dominate the local error, the currently employed defect estimates are unreliable. More research is required to develop effective strategies for detecting and coping with this situation.

References

[1] W.H. Enright, A new error control for initial value solvers, Appl. Math. Comp., 31, 1989, pp.288-301.

[2] W.H. Enright, Analysis of error control strategies for continuous Runge-Kutta methods, SIAM J. Numer. Anal., 26, 3, 1989, pp.588-599.

[3] W.H. Enright, The relative efficiency of alternative defect control schemes for high order Runge-Kutta formulas, SIAM J. Numer. Anal., 30, 5, 1993, pp. 1419-1445.

[4] W.H. Enright and H. Hayashi, A delay differential equation solver based on a continuous Runge-Kutta method with defect control, Numerical Algorithms, 16, 1997, pp. 349-364.

[5] W.H. Enright and H. Hayashi, The evaluation of numerical software for delay differential equations, in 'The quality of numerical software: assessment and enhancement', R. Boisvert, ed., 1997, Chapman and hall, London, pp.179-197.

[6] W.H. Enright and H. Hayashi, Convergence analysis of the solution of retarded and neutral delay differential equations by continuous methods, SIAM J. Numer. Anal., 35, 2, 1998, pp.572-585.

[7] W.H. Enright and M. Hu, Continuous Runge-Kutta methods for neutral Volterra integro-differential equations with delay, Applied Numerical Mathematics, 24, 1997, pp.1-16.

[8] W.H. Enright, K.R. Jackson, S.P. Nørsett and P.G. Thomsen, Interpolants for Runge-Kutta formulas, ACM Trans. Math. Soft., 12, 1986, pp.193-218.

[9] W.H. Enright, K.R. Jackson, S.P. Nørsett and P.G. Thomsen, Effective solution of discontinuous IVPs using a Runge-Kutta formula pair with interpolants, Appl. Math. Comput., 27, 1988, pp.313-335.

[10] W.H. Enright and P.H. Muir, A Runge-Kutta type boundary value ODE solver with defect control, SIAM J. Sci. Comp., 17, 1996, pp.479-497.

[11] W.H. Enright and P.H. Muir, Superconvergent interpolants for the collocation solution of BVODEs, to appear in SIAM J. Sci. Comp.

[12] W.H. Enright and J.D. Pryce, Two FORTRAN packages for assessing initial value methods, ACM Trans. Math. Soft., 13, 1, 1987, pp.1-27.

[13] W.H. Enright and R. Sivasothinathan, Superconvergent interpolants for collocation methods applied to mixed order BVODEs, submitted to ACM Trans. Math. Soft., 1999.

[14] I. Gladwell, L.F. Shampine, L.S. Baca and R.W. Brankin, Practical aspects of interpolation in Runge-Kutta codes, SIAM J. Sci. Statist. Comput., 8, 2, 1987, pp.322-341.

[15] D.J. Higham, Defect estimation in Adams PECE codes, SIAM J. Sci. Comp., 10, 1989, pp.964-976.

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

[17] D.J. Higham, Runge-Kutta defect control using Hermite-Birkhoff interpolation, SIAM J. Sci. Comp., 12, 1991, pp.991-999.

[18] M.K. Horn, Fourth- and fifth-order scaled Runge-Kutta algorithms for treating dense output, SIAM J. Numer. Anal., 20, 3, 1983, pp.558-568.

[19] T.E. Hull, The effectiveness of numerical methods for ordinary differential equations, SIAM Stud. Numer. Anal., 2, 1968, pp.114-121.

[20] J. Kierzenka and L.F. Shampine, A BVP solver based on residual control and the MATLAB PSE,

[21] C. MacDonald, A new approach for DAEs, Ph. D. Thesis, Dept. of Comp. Sc., Univ. of Toronto, 1999 (also appeared as DCS Tech. Rep. No. 317/99).

[22] H. Nguyen, Interpolation and error control schemes for algebraic differential equations using continuous implicit Runge-Kutta methods, Ph. D. Thesis, Dept. of Comp. Sc., Univ. of Toronto, 1995, (also appeared as DCS Tech. Rep. No. 298/95).

[23] S.P. Nørsett and G. Wanner, Perturbed collocation and Runge-Kutta methods, Numer. Math., 38, 1981, pp.193-208.

[24] L.F. Shampine, Interpolation for Runge-Kutta methods, SIAM J. Numer. Anal., 22, 5, 1985, pp.1014-1027.

[25] H.J Stetter, Cosiderations concerning a theory for ODE-solvers, in 'Numerical treatment of differential equations', R. Bulirsch, R.D. Grigorieff and J. Schroder, eds, lect notes in Mathematics 631, Spriger-Verlag, 1978, pp.188-200.

[26] H.J. Stetter, Interpolation and error estimation in Adams PC-codes, SIAM j¿ Numer. Anal., 16, 2, 1979, pp.311-323.

[27] J.H. Verner, Differentiable interpolants for high-order Runge-Kutta methods, SIAM J. Numer. Anal., 30, 5, 1993. pp.1446-1466.

[28] M. Zennaro, Natural continuous extensions of Runge-Kutta methods, Math. Comput., 46, 1986, pp.119-133.