The Reliability/Efficiency Trade-off for a New Class of ODE Solvers [a]

W.H. Enright

Department of Computer Science

University of Toronto

ICIAM 2007, Zurich

[a] This research was supported by the Natural Science and Engineering Research Council of Canada.

Acknowledgement

This work is part of an ongoing project and has benefited from numerous discussions and collaborations with

- Colleagues:
  - Paul Muir
  - Wayne Hayes
  - John Pryce
  - Ned Nedialkov
- Students:
  - Hossein Zivari Piran
  - Li Yan

Outline of Talk

An investigation of the Cost/Reliability trade-off in the numerical solution of ODEs.

- Continuous RK Methods for ODEs
- Defect Error Control for CRK Methods
- Relaxed and Strict Defect Control for CRKs
- Measuring Reliability and Verifying Validity
- Numerical Results
- Conclusions and Related Investigations

Continuous Runge-Kutta Methods

- Consider an IVP defined by the system

$$y' = f(x,y), \;\; y(x_0) = y_0, \;\; \text{on} \;\; [x_0, x_F].$$

- A numerical method will introduce a partitioning $x_0 < x_1 < \cdots < x_N = x_F$ and corresponding discrete approximations $y_0, y_1 \cdots y_N$. The y_i's are usually determined sequentially.

- On step $(i+1)$ let $z_i(x)$ be the solution of the local IVP:

$$z_i' = f(x, z_i(x)), \;\; z_i(x_i) = y_i, \;\; \text{on} \;\; [x_i, x_{i+1}].$$

CRK methods (cont)

A pth-order, s-stage, RK formula determines

$$y_{i+1} = y_i + h_{i+1} \sum_{j=1}^{s} \omega_j k_j,$$

where the jth stage is defined by,

$$k_j = f(x_i + h_{i+1} c_j, y_i + h_{i+1} \sum_{r=1}^{s} a_{jr} k_r).$$

A Continuous extension (CRK) is determined by adding $(\bar{s} - s)$ extra stages to obtain an order p approximation for $x \in (x_i, x_{i+1})$

$$u_i(x) = y_i + h_{i+1} \sum_{j=1}^{\bar{s}} b_j(\frac{x - x_i}{h_{i+1}}) k_j,$$

where $b_j(\tau)$ is a polynomial of degree at least p and $\tau = \frac{x - x_i}{h_{i+1}}$.

CRK methods (cont)

- We will consider two types of $O(h^p)$ extensions, satisfying:

$$u_i(x) = y_i + h_{i+1} \sum_{j=1}^{\bar{s}} b_j(\tau) k_j = z_i(x) + O(h^{p+1}).$$

- The $[u_i(x)]_{i=1}^N$ define a piecewise polynomial $U(x)$ for $x \in [x_0, x_F]$. This can be considered the numerical solution generated by the CRK method.
- $U(x) \in C^0[x_0, x_F]$ and will interpolate the underlying discrete RK values, y_i, if $b_j(1) = \omega_j$ for $j = 1, 2 \cdots s$ and $b_{s+1}(1) = b_{s+2}(1) = \cdots b_{\bar{s}}(1) = 0$.
- Similarly a simple set of constraints on the $b_j^{'}(\tau)$, will ensure $U'(x)$ interpolates $f(x_i, y_i)$ and therefore $U(x) \in C^1[x_0, x_F]$.

Defect Error Control

$U(x)$, the numerical solution of the ODE, has an associated defect,

$$\delta(x) = f(x, U(x)) - U'(x) = f(x, u_i(x)) - u_i'(x), \quad \text{for} \quad x \in [x_i, x_{i+1}].$$

It can be shown that, for such a CRK,

$$\delta(x) = G(\tau) h_{i+1}^{p} + O(h_{i+1}^{p+1}),$$

$$G(\tau) = q_1(\tau) F_1 + q_2(\tau) F_2 + \cdots + q_k(\tau) F_k,$$

where the $q_j s$ are polynomials in τ that depend only on the method while the $F_j s$ are constants that depend only on the problem.

Methods can be implemented to adjust h_{i+1} in an attempt to ensure that the maximum magnitude of $\delta(x)$ is bounded by TOL on each step.

Defect Error Control (cont)

$$\delta(x) = G(\tau)h_{i+1}^{p} + O(h_{i+1}^{p+1}),$$

$$G(\tau) = q_1(\tau)F_1 + q_2(\tau)F_2 + \cdots + q_k(\tau)F_k.$$

- As $h \rightarrow 0$ the defect will then look like a linear combination of the $q_j(\tau)$ over $[x_i, x_{i+1}]$.
- In the special case where $k = 1$ the shape of the defect will be the same (as $h \rightarrow 0$) for all problems and all steps. That is, the defect will almost always 'converge' to a multiple of $q_1(\tau)$.

Defect Error Control (cont)

- When $k > 1$ one can estimate the maximum defect by evaluating $\delta(x)$ at a carefully chosen set of sample points. We will call this defect control strategy, **Relaxed Defect Control (RDC)**.
- The maximum defect will be easier to estimate if $k = 1$, in which case the maximum should occur (as $h \to 0$) at $\tau = \tau^*$ where τ^*, the location in $[0, 1]$ of the local maximum of $q_1(\tau)$. In this case we will refer to the defect control stategy as **Strict Defect Control (SDC)**.
- We will consider 2 types of continuous extensions: $u_i(x)$ corresponding to RDC and $\tilde{u}_i(x)$ corresponding to SDC.

Defect Control (cont)

$$RDC: \quad u_i(x) = y_i + h_{i+1} \sum_{j=1}^{\bar{s}} b_j(\tau) k_j = z_i(x) + O(h^{p+1}),$$

$$SDC: \quad \tilde{u}_i(x) = y_i + h_{i+1} \sum_{j=1}^{\tilde{s}} \tilde{b}_j(\tau) k_j = z_i(x) + O(h^{p+1}).$$

Formula	p	s	$\bar{s}$	$\tilde{s}$
CRK4	4	4	6	8
CRK5	5	6	9	12
CRK6	6	7	11	15
CRK7	7	9	15	20
CRK8	8	13	21	27

Table 1: Cost per step of some CRK formulas

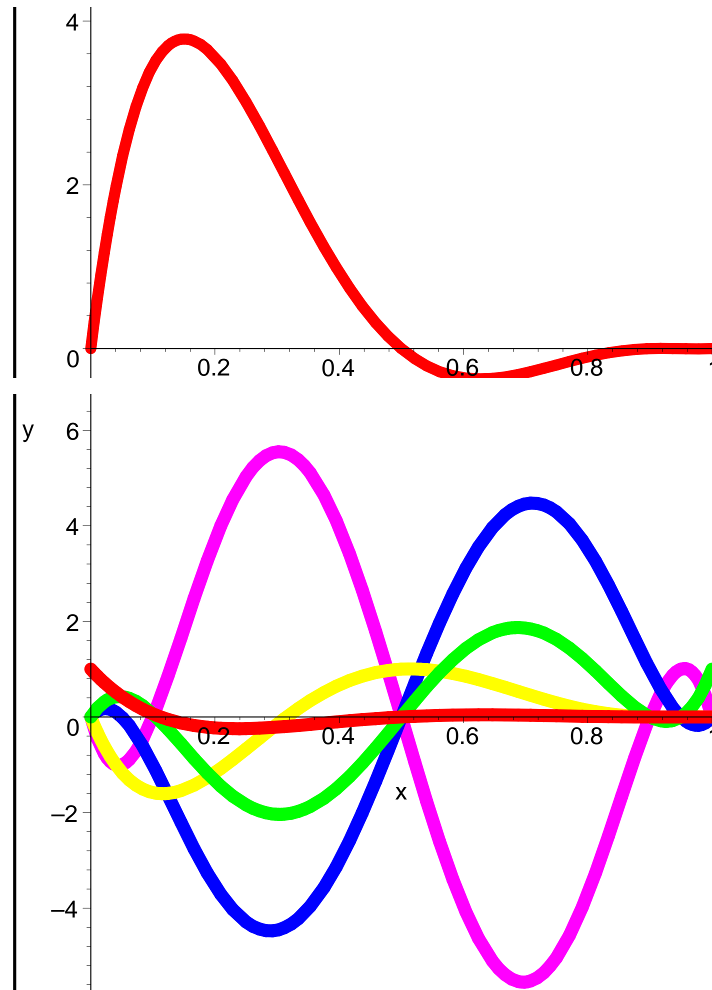
Strict Defect Control

$$\delta(x) = q_1(\tau) F_1 h_i^p \;+\; (\hat{q}_1(\tau)\hat{F}_1 + \hat{q}_2(\tau)\hat{F}_2 + \cdots\cdots \hat{q}_k(\tau)\hat{F}_k) h^{p+1} + O(h_i^{p+2})$$

Potential Difficulties:

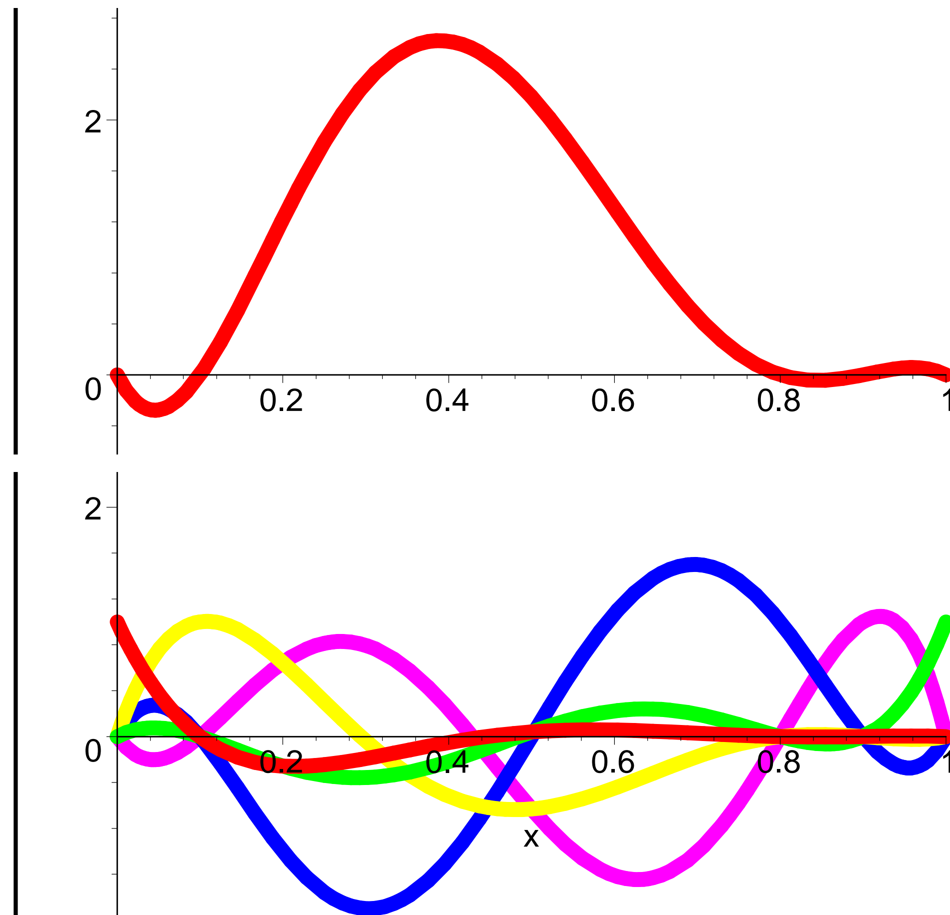
- $q_1(\tau)$ may have a large maximum (its 'average' value must be one).
- The $\hat{q}_j(\tau)$ may be large in magnitude relative to $q_1(\tau)$ (and therefore h would have to be small before the estimate is justified. (That is, before $|h\hat{q}_j(\tau)| << |q_1(\tau)|$.)
- $|F_1|$ may be zero (or very small) on isolated steps.

Strict Defect Control: Ex. I



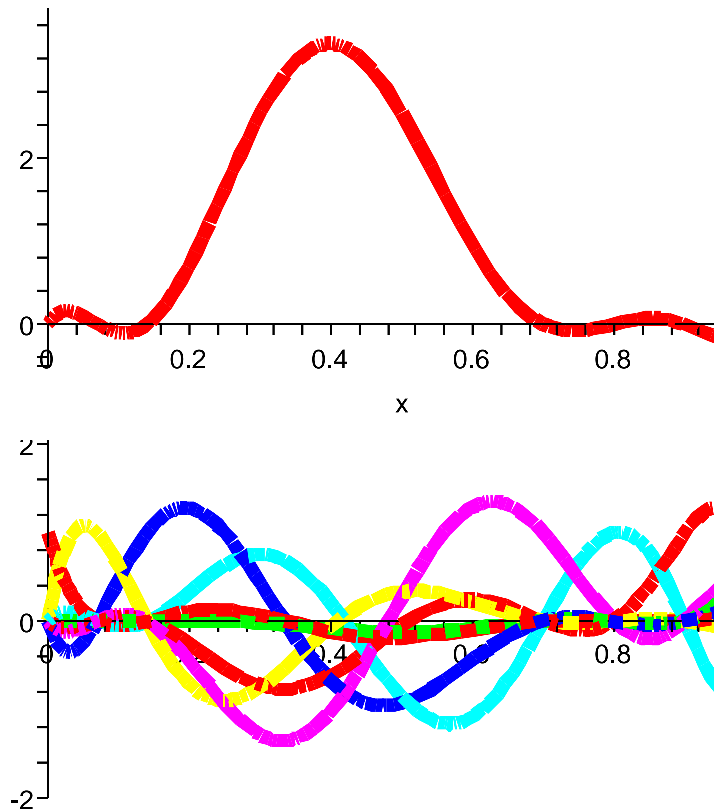
Graphs of $q_1(\tau)$ (top) and selected $\hat{q}_j(\tau)$ for a typical CRK6.

Strict Defect Control: Ex. II



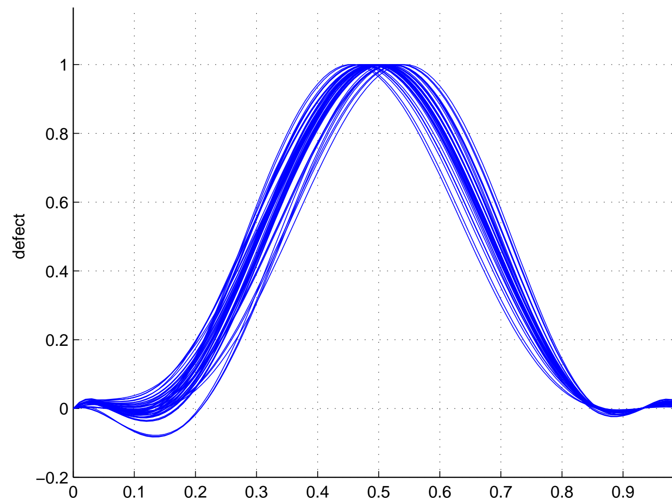
$q_1(\tau)$ (top) and selected $\hat{q}_j(\tau)$ for a 'better' (near optimal) CRK6.

Strict Defect Control: Ex. III



Graphs of $q_1(\tau)$ (top) and selected $\hat{q}_j(\tau)$ for a near-optimal interpolant for CRK8.

Typical Shape of SDC Defects



Plot of defect/τ for each step required to solve a typical problem with CRK6 and $TOL = 10^{-6}$.

Quantifying Reliability

Consider two measures of reliability:

- How well does a **Method** control the maximum magnitude of the defect? We can measure the ratio of the max defect to TOL and the fraction of steps where this ratio is greater than 1 ?
- How well does the **Estimate** of the max defect reflect its true value? We can measure both the ratio of the true maximum defect (on a step) to the estimated value and the fraction of steps where the estimated maximum is within one percent of the true maximum.

Improving the Defect Estimate

- Extensive testing of CRKs with SDC revealed that the few steps where the estimate was too small were inevitably associated with steps where $|F_1|$ was near zero. On these isolated steps, the actual max defect was usually smaller than TOL, but the estimated value was too small.
- We introduced a validity check to detect such difficulties. If this check is satisfied on all steps, one can have increased confidence in the reliability of the integration.
- We have modified the basic SDC strategy so that, when this check is not satisfied, 2 extra sampled defect evaluations are performed to determine a more suitable defect estimate. We call this modified strategy SDCV.

The Validity Check

- We know $q_1(\tau)$ is zero at $\tau = 0, \tau = 1$ and attains its maximum magnitude at τ^*. Let $\tau_1 < \tau^*$ and $\tau_2 > \tau^*$ satisfy

$$q_1(\tau_1) = q_1(\tau_2) = q_1(\tau^*)/2.$$

- If we then let $R_1 = \frac{\delta(x_i+\tau_1 h)}{\delta(x_i+\tau^* h)}$ and $R_2 = \frac{\delta(x_i+\tau_2 h)}{\delta(x_i+\tau^* h)}$, then as $h \to 0$ we expect R_1 and R_2 to approach $1/2$. We compute these two ratios and consider the validity check to be satisfied if both are close to $1/2$.

- In our tests we interpreted 'close to $1/2$' to mean 'in the range $[.3, .7]$'. For our formulas:
  CRK5: $\tau^* = .389, \tau_1 = .207, \tau_2 = .600.$
  CRK6: $\tau^* = .500, \tau_1 = .311, \tau_2 = .689.$
  CRK8: $\tau^* = .500, \tau_1 = .353, \tau_2 = .647.$

Implementation and Testing

- We have implemented 3 versions of CRK5, CRK6 and CRK8: RDC, SDC and SDCV. The user selects the defect control strategy by setting an integer parameter.
- We have run all versions on the 25 test problems of DETEST (all non-stiff), at 9 tolerances from 10^{-1} to 10^{-9}.
- We report summaries only, but detailed results are available. We report two measures of efficiency: NSTP and NFCN, two measures of the reliability of the method : DMAX and Frac-D (max defect and fraction of steps where this exceeded TOL), and two measures of the reliability of the estimate: R-Max and Frac-G (maximum ratio of the true maximum defect to the estimate and the fraction of steps where this was bounded by 1.01).

Numerical Results for CRK5X

Results on the 25 DETEST Problems for an order 5 CRKX with 3 Defect Control Strategies
(The underlying discrete RK formula is the same as ode45 of Matlab)

TOL	CRK	NSTP	NFCN	DMAX	Frac-D	R-Max	Frac-G
10^{-2}	RDC	609	7153	2.37	.199	18.85	.18
	SDC	623	9853	1.02	.003	8.12	.63
	SDCV	625	11709	0.97	.000	1.05	.67
10^{-4}	RDC	1070	12130	5.89	.179	126.82	.14
	SDC	1065	16081	1.60	.005	7.12	.73
	SDCV	1065	19033	1.01	.001	1.12	.78
10^{-6}	RDC	2176	23146	5.44	.233	55.44	.09
	SDC	2095	30037	1.44	.007	11.49	.83
	SDCV	2099	35703	1.01	.002	1.08	.86
10^{-8}	RDC	4929	46051	21.28	.354	207.40	.07
	SDC	4562	56953	1.24	.003	32.80	.94
	SDCV	4566	66937	1.01	.001	1.07	.95

Numerical Results for CRK6X

Results on the 25 DETEST Problems for an order 6 CRK with 3 Defect Control Strategies

TOL	CRK	NSTP	NFCN	DMAX	Frac-D	R-Max	Frac-G
10^{-2}	RDC	552	7879	5.27	.176	23.25	.50
	SDC	547	10585	1.00	.000	1.74	.70
	SDCV	549	12300	1.00	.000	1.43	.71
10^{-4}	RDC	955	13082	4.87	.144	15.34	.55
	SDC	929	17305	4.90	. 003	18.90	.87
	SDCV	931	19819	1.00	.001	1.08	.87
10^{-6}	RDC	1789	23499	10.75	.103	112.90	.59
	SDC	1748	30925	1.01	.001	1.81	.96
	SDCV	1748	35073	1.01	.001	1.08	.96
10^{-8}	RDC	3622	43288	6.48	.098	1286.90	.67
	SDC	3547	57460	1.01	.001	1.14	.98
	SDCV	3547	65148	1.01	.001	1.07	.98

Numerical Results for CRK8X

Results on the 25 DETEST Problems for an order 8 CRK with 3 Defect Control Strategies

TOL	CRK	NSTP	NFCN	DMAX	Frac-D	R-Max	Frac-G
10^{-2}	RDC	337	8745	10.68	.213	36.71	.30
	SDC	332	11439	7.16	.009	30.50	.35
	SDCV	333	12793	1.01	.003	1.65	.35
10^{-4}	RDC	495	13285	7.70	.139	32.70	.17
	SDC	466	15781	1.02	. 002	4.34	.45
	SDCV	465	17319	1.05	.004	1.47	.45
10^{-6}	RDC	715	18245	6.09	.126	134.32	.10
	SDC	707	23425	3.01	.008	22.70	.58
	SDCV	712	26253	1.02	.001	1.34	.59
10^{-8}	RDC	1095	27065	31.12	.179	409.09	.08
	SDC	1081	34787	1.86	.005	20.80	.62
	SDCV	1081	38251	1.12	.007	2.60	.62

Observations/Conclusions

- RDC delivers acceptable reliability (for most applications) at all tolerances.
- SDC without the validity check is usually very reliable at all tolerances. The extra cost (over RDC) is around 25 %.
- SDC with a validity check is very reliable at all tolerances and will signal those steps where the defect may be slightly larger than expected. The extra cost (over RDC) is around 50%.

Future/Ongoing Investigations

- Consider the use of CRKs with SDC in BV solvers, in DAE solvers and in DDE solvers.
- Extension of approach to variable order Adams methods for IVPs.
- Extension of approach to stiff solvers such as those based on BDF or IRK formulas.