On low order embedded pairs of implicit Runge-Kutta formulas

Frank Cameron and Robert Piché
Tampere University of Technology
P.O. Box 692, 33101 Tampere, Finland
email: frank@helium.ee.tut.fi, piche@cc.tut.fi

Abstract

Implicit Runge-Kutta methods are used for solving stiff ODEs such as those arising in mechanical or electrical system simulation and in semidiscretisation of partial differential equation evolution problems. Embedding one Runge-Kutta formula with another is a way of obtaining an estimate of the local error (for step size control) at a modest computation cost. Our interest is with the design of embedded pairs of low order. We consider both accuracy and basic stability properties of Runge-Kutta formulas with an eye to the performance of the pair as a whole. We present some negative results showing that embedded pairs with certain combinations of stability properties cannot exist. Finally we analyze and compare 7 pairs from the literature and 6 new pairs.

1 Introduction

Implicit Runge-Kutta formulas of relatively low order (order 2 or 3), such as the trapezoid rule, are often used for solving stiff ordinary differential equations arising in mechanical or electrical system simulation and in semidiscretisation (method of lines, Rothe method) of partial differential equation evolution problems.

Although implicit Runge-Kutta formulas have been studied for many years, new classes of methods continue to appear. Hosea and Shampine [8] have recently studied low order formulas with the novel property that they contain both implicit and explicit stages. They present several negative results asserting the nonexistence of formulas with certain combinations of stability and accuracy properties. Such negative results are obviously helpful in directing the search for new formulas, and they complement the negative results gathered in the survey paper by Alexander [1] and in the monograph of Hairer and Wanner [5].

The purpose of having an embedded pair of Runge-Kutta formulas is to get an estimate of local error. This estimate is typically used for controlling the step size used when advancing the solution. In this paper we try to keep in mind the interaction between the two formulas in the pair and what this interaction implies for the operation of the pair as a whole. For example, it would not make sense to pack one formula with desirable properties if the consequence was that the other formula suffered to such a degree that the embedded pair performed poorly.

We include results for pairs whose implicit parts are based either on singly implicit (SIRK) or singly diagonally implicit (SDIRK) Runge-Kutta formulas.

After presenting some definitions in section 2, we review in section 3 some relevant stability properties and discuss what the stability properties imply for the behavior of the pair. Section 4 contains a number of negative results asserting the nonexistence of pairs with certain combinations of properties. We consider accuracy properties in section 5, again focusing on what is desirable so that the pair perform well. The last section contains analysis and comparison of seven pairs from the literature and six new pairs.

2 Definitions

The Butcher table for a pair of embedded Runge-Kutta (RK) formulas is given as follows:

$$\begin{array}{c|c} c & A \\ \hline & b^T \\ \hline & \hat{b}^T \end{array} \tag{1}$$

In this table A is an $s \times s$ matrix and $c = Ae$, where e is a vector of ones. The vectors b and $\hat{b}$ are associated with the lower and higher order RK methods respectively; each has s elements. We will consider methods involving both implicit and explicit stages.

The RK formula associated with b will be called the *estimator formula* or the *b-formula*. The *auxiliary formula* or the *$\hat{b}$-formula* is that associated with $\hat{b}$. Unless otherwise stated, we will assume that the estimator formula is used to advance

the state. The orders of the b-formula and the $\hat{b}$-formula are denoted by p and $\hat{p}$ respectively.

We will be considering methods whose A matrix is block-lower diagonal with the following structure:

$$A = \begin{bmatrix} A_1 & 0 & 0 \\ A_2 & A_3 & 0 \\ A_4 & A_5 & A_6 \end{bmatrix} . \tag{2}$$

In this partition, A_1 and A_6 are lower triangular matrices whose diagonal elements are zero and A_3 is an $s_i \times s_i$ matrix whose rank is s_i. The number of implicit and explicit stages in an RK formula are denoted by s_i and $s_e \equiv (s - s_i)$ respectively.

An *Implicit RK formula* (IRK) has $s_i > 0$. A *Diagonally Implict RK formula* (DIRK) is one where A_3 in (2) is lower triangular. A *Singly Diagonally Implicit RK formula* (SDIRK) is a DIRK formula where all diagonal elements of A_3 in (2) are the same. A *Singly Implicit RK formula* (SIRK) is one where A_3 has one real s_i-fold eigenvalue, μ, and A_3 is not lower triangular. As defined here, SIRK formulas and SDIRK formulas are mutually exclusive.

Next we will define the stability functions we will use. We assume we are starting from y_n corresponding to time t_n. We restrict ourselves to a linear problem of the form

$$y' = \lambda y + q \ , \quad \lambda \in \mathcal{C} \ . \tag{3}$$

We will assume $p < \hat{p}$. The numerical estimate to the true solution at $t_{n+1} \equiv t_n + h$, provided by the lower order estimator formula satisfies

$$y_{n+1} = R(h\lambda)(y_n - q) + q \ . \tag{4}$$

The stability function for the estimator formula is given by

$$R(z) \equiv 1 + z b^T (I - zA)^{-1} e \tag{5}$$

where $z = h\lambda$ and $e^T = [1, 1, \ldots 1]$. Similarly the higher order auxilary formula is given by

$$\hat{y}_{n+1} = \hat{R}(h\lambda)(y_n - q) + q \ , \tag{6}$$

where

$$\hat{R}(z) \equiv 1 + z \hat{b}^T (I - zA)^{-1} e \tag{7}$$

The error estimate, δ, is calculated from

$$\delta_{n+1} \equiv \hat{y}_{n+1} - y_{n+1} = R_\delta(z) \, (y_n - q) \ . \tag{8}$$

where the error estimate stability function is given by

$$R_\delta(z) \equiv \hat{R}(z) - R(z) \ . \tag{9}$$

We will have use for the following limits:

$$\gamma \equiv \lim_{|z| \to \infty} R(z) \ , \tag{10}$$

$$\hat{\gamma} \equiv \lim_{|z|\to\infty} \hat{R}(z) \ . \tag{11}$$

A stability function is a rational function. An RK formula with stability function $R(z) = G(z)/H(z)$ is *proper* if degree$(G) \leq$ degree(H) or equivalently if $\gamma < \infty$. An RK formula with stability function, $R(z) = G(z)/H(z)$ is *strictly proper* if degree$(G) <$ degree(H) or equivalently if $\gamma = 0$.

3 Runge-Kutta formula stability properties

Many different stability properties have been developed for RK formulas. Hairer and Wanner [5] provide a detailed and comprehensive survey. We will focus our attention on what certain basic stability properties imply for the embedded pair as a whole.

When dealing with stiff ODE's we must concern ourselves with the operation of the embedded pair for both large $|z|$ and small $|z|$, where $z = \lambda h$ (see (5)). Stability issues arise for large $|z|$, so we will be concerned with large $|z|$ in this section.

Properness Properness was defined in section 2. The significance of properness can be seen by studying the test problem (3) with constant q. Given the initial conditions $y(t_n) = y_n$ the exact solution to this test problem at time $t_{n+1} = t_n + h$ is given by y^*:

$$y^*(t_{n+1}) = q + \exp(h\lambda)(y_n - q) \ .$$

The local error of the estimator formula (4) is hence

$$\varepsilon_{n+1} \equiv y^*(t_{n+1}) - y_{n+1} = (\ \exp(h\lambda) - R(h\lambda)\ )(y_n - q) \ . \tag{12}$$

We will assume that $h > 0$ and that the linear system (3) is stable, i. e. $\mathrm{Re}(\lambda) < 0$. Using $z = h\lambda$, we can find the limit of the local error:

$$\lim_{|z|\to\infty} |\varepsilon_{n+1}| \ = \ \lim_{|z|\to\infty} |\,(\ \exp(z) - R(z)\ )(y_n - q)\,| \ = \ \gamma\,|y_n - q| \ . \tag{13}$$

If $R(z)$ is not proper, the local error $|\varepsilon_{n+1}|$ will grow as $|z|$ grows. Clearly properness is important for the estimator formula. Now consider the auxiliary function. We will assume $\gamma < \infty$. Using (10) and (11), the limit of the error estimate from (8) can be written as follows:

$$\begin{aligned} \lim_{|z|\to\infty} |\delta_{n+1}| \ &= \ \lim_{|z|\to\infty} \left|\,(\ \hat{R}(z) - R(z)\ )(y_n - q)\,\right| \\ &= \ |\,\hat{\gamma} - \gamma\,|\ |y_n - q| \end{aligned} \tag{14}$$

If $\hat{R}$ is not proper, then $\hat{\gamma} \to \infty$ as $|z| \to \infty$. Hence although the size of the local error from (12) is actually $\gamma|y_n - q|$, the estimated error grows without bound as $|z|$ increases. If $\hat{R}$ is not proper we can thus expect the local error to be overestimated, sometimes grossly. This will make any step size adjustment overly conservative and this will cost us in the number of steps needed to integrate to a certain point. Now let us consider the situation where both $\hat{R}(z)$ and $R(z)$ are proper but not strictly

proper. Given $\gamma > 0$, we want to know what is a desirable value for $\hat{\gamma}$. From (14) and (13) it follows that if $|\hat{\gamma} - \gamma| < |\gamma|$, then for large $|z|$ the local error will be underestimated. This can occur for $\hat{\gamma} \approx \gamma$. Underestimation of error is to be avoided. For an exact error estimate for large $|z|$ we would need $|\hat{\gamma} - \gamma| = |\gamma|$, i. e. $\hat{\gamma} = 0$ or $\hat{\gamma} = 2\gamma$. We do not suggest that one *must* have $\hat{\gamma} = 0$ or $\hat{\gamma} = 2\gamma$, since this result comes from setting $|z| = \infty$, which is impossible. Still we do suggest avoiding $\hat{\gamma} \approx \gamma$ when $\gamma \neq 0$.

A-stability An RK formula for which $|R(z)| \leq 1$ when $\mathrm{Re}(z) < 0$ is A-stable. Let us consider again the test problem of (3). It is important for the estimator formula (4) to be A-stable since we are using the output of (4) to advance y. If $R(z)$ is A-stable and (3) is stable, then we may vary h arbitrarily without having to worry about the stability of the sequence $[\, y_{n+1}, y_{n+2}, \ldots \,]$ that is produced by (4). The estimator formula retains or inherits the stability of (3). Consider now the auxiliary formula (6). We are only using (6) one step at a time, i. e. the initial condition used in (6) is a previous estimate y_n provided by the estimator formula (4). In contrast to properness, we cannot see any obvious reason to require the auxiliary formula to be A-stable.

L-stability An RK formula is L-stable if it is both A-stable and strictly proper, i. e. $\gamma = 0$ (see 10). Using the definitions of section 2 we obtain

$$\text{L-stable formula} \;\Rightarrow\; \text{strictly proper formula} \tag{15}$$

$$\text{L-stable formula} \;\Rightarrow\; \text{A-stable formula} \;\Rightarrow\; \text{proper formula} \tag{16}$$

We continue the analysis presented above. Consider the local error limit in (13). When the estimator formula is L-stable, γ is zero, and hence the local error tends to 0. One should approach this seemingly good news with some caution. Although $|\varepsilon_{n+1}| \to 0$ as $|z|$ grows, $|\varepsilon_{n+1}|$ does not necessarily tend to zero with as much accuracy as one might expect. Prothero and Robinson [10] have shown that the apparent order of many RK methods decrease significantly as $|z|$ grows. Let us consider how the estimate of the local error behaves. If the estimator formula is L-stable and the auxiliary formula is not, then $\hat{\gamma} > 0$ and $\gamma = 0$. It follows that the local error estimate in (14) will be too large since the true local error limit in (13) tends to zero. This in turn implies overly conservative step size adjustment. Note though, that the value of $\hat{\gamma}$ does matter: the overestimation of local error for large $|z|$ will be less for smaller $\hat{\gamma}$ values.

S-stability The concept of S-stability was introduced by Prothero and Robinson [10]. They had noticed that A-stability was not enough to ensure stability for large stiff problems. Their analysis was based on the following test problem

$$y' = g'(t) + \lambda(y - g(t)) \;, \quad \lambda \in \mathcal{C} \;. \tag{17}$$

They show that the solution provided by a Runge-Kutta method can be written in the following form:

$$y_{n+1} = R(z)\,(\, y_n - g(t_n)\,) + g(t_{n+1}) + h\beta(\, z, \mathbf{g}, \mathbf{g}', g(t_n), g(t_{n+1})\,) \;, \tag{18}$$

where $z = h\lambda$, $\mathbf{g}$ and $\mathbf{g}'$ are vectors representing the evaluation of g and g' at the points determined by the c of (1). (Note that the Prothero and Robinson's z is the inverse of the z used here.) For (17) the true solution is $y^*(t) = g(t)$, hence the local error of (18) is

$$\varepsilon_{n+1} \equiv y^*(t_{n+1}) - y_{n+1} = R(z)(y_n - g(t_n)) + h\beta(z) \ , \tag{19}$$

where for simplicity we have dropped all the arguments on β except for z. If the Runge-Kutta method is *S-stable*, it essentially means that for finite real h, the $|h\beta(z)|$ term in (19) is bounded for $|z| \to \infty$ and $\mathrm{Re}(\lambda) < 0$. A *strongly S-stable* method has $|\beta(z)| \to 0$ under the same conditions. Clearly, strong S-stability would appear to be desirable for the estimator formula when dealing with stiff problems. Let us consider what S-stability or strong S-stability implies for the local error estimate. The local error estimate for (17) will be

$$\delta_{n+1} = (\hat{R}(z) - R(z)) (y_n - g(t_n)) \ + \ h(\hat{\beta}(z) - \beta(z)) \ . \tag{20}$$

The $(\hat{R}(z) - R(z))$ term we have discussed above when we considered properness and L-stability. The $(\hat{\beta}(z) - \beta(z))$ term is new but the conclusions are familiar. If the estimator formula is S-stable (or strongly S-stable) and the auxiliary formula is not, we can expect overestimation of error and the step sizes selected will be conservative. Finally to relate these S-stability properties to those stability properties previously discussed, we note that for a formula to be S-stable it must be A-stable, and for a formula to be strongly S-stable it must be L-stable.

4 Some negative results

In this section we present results showing that embedded pairs of IRK formulas with certain properties do *not* exist. The purpose of these negative results is to eliminate certain possibilities when one is designing embedded IRK formulas.

Table 1 summarizes the results of this section. To interpret the results in this table one can use (15), (16) and the definitions provided at the end of sec. 2. For example, from (15) it follows that a formula must be strictly proper for it to be L-stable.

We start with a result that shows that we will need at least 3 implicit stages to get an embedded pair with orders $(p, \hat{p}) = (2, 3)$ where the b-formula is strictly proper and the $\hat{b}$-formula is proper.

Theorem 1 *No pair of embedded SDIRK formulas exists with the following properties: (a) $p = 2$ and the b-formula is stricly proper, (b) $\hat{p} = 3$ and the $\hat{b}$-formula is proper, and (c) $s_i < 3$.*

Proof The case of $s_i = 1$ is trivial and we shall not include it. Let us consider $s_i = 2$. Since we are dealing with SDIRK formulas the diagonal elements of A_3 of (2) are all the same. Let this diagonal element be μ. We will show that the conditions imposed result in two incompatible conditions on μ.

Table 1: A summary of the results of section 4 regarding different types of Runge-Kutta embedded pairs. "Stricly proper" has been shortened to st. proper (see sec. 2).

type	no. of stages	(property, order) for		conclusion	thm.
		b-formula	$\hat{b}$-formula		
SDIRK	$s_i < 3$, $s_e \geq 0$	(st. proper,2)	(proper,3)	no pair exists	1
SIRK	$s_i < 3$, $s_e \geq 0$	(st. proper,2)	(proper,3)	no pair exists	2
SDIRK	$s_i < 4$, $s_e \geq 0$	(st. proper,3)	(proper,4)	no pair exists	3
SDIRK	$s_i = 3$, $s_e = 0, 1$	(st. proper,2)	(st. proper,3)	no pair exists	4, 5
SIRK	$s_i = 3$, $s_e = 0$	(st. proper,2)	(st. proper,3)	no pair exists	6, 7
IRK	$s_i = 3$, $s_e \geq 0$	(st. proper,2)	(st. proper,3)	$R(z) = \hat{R}(z)$	9

The stability function for the b-formula is given by (5). Let $G(z)$ and $H(z)$ be the numerator and denominator polynomials of $R(z)$:

$$R(z) \;=\; \frac{G(z)}{H(z)} \;=\; \frac{\sum_{j=0}^{s} g_j\, z^j}{\sum_{j=0}^{s} h_j\, z^j} \tag{21}$$

From (5) we may write

$$R(z) = 1 + \frac{z b^T G_A(z) e}{\det(I - zA)}\ , \tag{22}$$

where

$$(I - zA)^{-1} = G_A(z) \,/\, \det(I - zA)\ , \tag{23}$$

and $G_A(z)$ is a matrix polynomial in A. Note that $H(z) = \det(I - zA)$. The $G_A(z)$ matrix polynomial can be expressed as

$$G_A(z) \;=\; \sum_{j=0}^{s} G_j\, z^j \tag{24}$$

where G_j can be obtained by expanding Leverrier's algorithm [13, pp. 117–8]:

$$G_j \;=\; \sum_{k=0}^{j} h_{j-k}\, A^k \tag{25}$$

It follows from (21), (22) and (24) that

$$g_j = h_j + b^T G_{j-1} e\ , \quad G_{-1} = 0\ . \tag{26}$$

The expressions (21)–(25) are valid for $\hat{R}(z)$ if we substitute $\hat{b}$ for b. Hence following the same development for the auxiliary formula, $\hat{R}(z)$, we can obtain

$$\hat{g}_j = h_j + \hat{b}^T G_{j-1} e\ , \quad G_{-1} = 0\ . \tag{27}$$

We will derive an expression for g_2, the coefficient of z^2 in the $G(z)$ polynomial. Using (26), the g_2 coefficient can be written as

$$g_2 = h_2 + b^T G_1 e\ . \tag{28}$$

From the definition of an SDIRK with $s_i = 2$, it follows that $\det(I - zA) = H(z) = (1 - \mu z)^2$. Using this and (25) we may write

$$g_2 = \mu^2 + b^T(-2\mu I + A)e \; . \tag{29}$$

Now the b-formula was of order 2, so the following two conditions must hold:

$$b^T e = 1 \; , \tag{30}$$

$$b^T c = b^T A e = 1/2 \; . \tag{31}$$

Using (30) and (31), in (29) we obtain

$$g_2 = \mu^2 - 2\mu + 1/2 \; . \tag{32}$$

We required the b-formula to be strictly proper. Since degree$(H) = 2$, for the b-formula to be strictly proper we must have $g_2 = 0$.

We now consider the $\hat{b}$-formula. In particular we want the coefficient of z^3 in the numerator of $\hat{R}(z)$. From (27) the coefficient of z^3 is given by

$$\hat{g}_3 \; = \; h_3 \; + \; \hat{b}^T G_2 \, e \; = \; h_3 \; + \; \hat{b}^T \left[\, h_2 I \; + \; h_1 A \; + \; h_0 A^2 \, \right] \; e \tag{33}$$

The $\hat{b}$-formula was of order 3 which means that the following four conditions hold:

$$\hat{b}^T e = 1 \; , \tag{34}$$

$$\hat{b}^T c = 1/2 \; , \tag{35}$$

$$\hat{b}^T \, (c \odot c) = 1/3 \; , \tag{36}$$

$$2\hat{b}^T A c = \hat{b}^T \, (c \odot c) \; , \tag{37}$$

where $c = Ae$ and $\odot$ represents the componentwise product, $(x \odot y)_i \equiv (x_i y_i)$. Using $H(z) = (1 - \mu z)^2$ and (34)—(37) in (33), we obtain

$$\hat{g}_3 \; = \; \mu^2 - 2\mu \hat{b}^T c + \hat{b}^T A c \; = \; \mu^2 - \mu + 1/6 \; . \tag{38}$$

From the condition that the $\hat{b}$-formula be proper and degree$(H) = 2$, it follows that $\hat{g}_3 = 0$. From above we also needed $g_2 = 0$. The roots of (38) and those of (32) do not intersect. ■

Remarks

Using almost the same arguments as used in the proof of Theorem 1 it can be shown that there does not exist an SDIRK formula that is proper with $p = 3$, $s_i = 1$ and $s_e > 1$. Similary it can be shown that there does not exist an SDIRK formula that is stricly proper with $p = 3$, $s_i = 2$ and $s_e > 1$.

We will next prove a similar result for an embedded SIRK pair (see sec. 2).

Theorem 2 *No pair of embedded SIRK formulas exists with the following properties: (a) its A matrix is given by (2), (b) $p = 2$ and the b-formula is stricly proper, (c) $\hat{p} = 3$ and the $\hat{b}$-formula is proper, and (d) $s_i < 3$.*

Proof We shall only present what is necessary and then refer to the proof of Theorem 1. We ignore the trivial case of $s_i = 1$ and consider only $s_i = 2$.

The stability function for the b-formula is given by (22). Since A_1 and A_6 in (2) are triangular with zero diagonals we can write

$$\det(I - zA) \;=\; \det(I - zA_1)\;\det(I - zA_3)\;\det(I - zA_6) \;=\; \det(I - zA_3) \qquad (39)$$

From the definition of a SIRK we know that $\det(I - zA_3) = (1-\mu z)^2$. The remainder of the proof is analogous to that of Theorem 1. ■

An analogous result to Theorem 1 exists for SDIRK pairs of orders 3 and 4.

Theorem 3 *No pair of embedded SDIRK formulas exists with the following properties: (a) $p = 3$ and the b-formula is stricly proper, (b) $\hat{p} = 4$ and the $\hat{b}$-formula is proper, and (c) $s_i < 4$.*

Proof The proof is analogous to the proof of Theorem 1, so we provide only those points that differ significantly from it. We consider only $s_i = 3$. We will derive expressions for g_3 of $R(z)$ and $\hat{g}_4$ of $\hat{R}(z)$ (see (21)). Using g_3 and $\hat{g}_4$ we will show that the requirements on the embedded pair result in incompatible conditions on μ, where all diagonal elements of A_3 of (2) are equal to μ.

From (26) we may write an expression for g_3:

$$g_3 \;=\; h_3 \;+\; b^T G_2 e\,. \qquad (40)$$

This expression is the same as (33) except that b replaces $\hat{b}$. For an SDIRK with $s_i = 3$, $\det(I - zA) = H(z) = (1 - \mu z)^3$. Taking this into account and using the order conditions (34)—(37) with $\hat{b}$ replaced by b, we can rewrite (40) as

$$g_3 \;=\; -\mu^3 + 3\mu^2 - 3\mu/2 + 1/6\,. \qquad (41)$$

The $\hat{b}$-formula was of order 4. In addition to conditions (34)—(37) holding, the following must be satisfied:

$$\hat{b}^T\,(c \odot c \odot c) = 1/4\,, \qquad (42)$$

$$3\hat{b}^T A(c \odot c) = \hat{b}^T\,(c \odot c \odot c)\,, \qquad (43)$$

$$\hat{b}^T A(c \odot c) = 2\,\hat{b}^T\,A^2 c\,. \qquad (44)$$

From (27) we may write an expression for $\hat{g}_4$:

$$\hat{g}_4 \;=\; h_4 \;+\; \hat{b}^T G_3 e\,. \qquad (45)$$

Using (25) and $H(z) = (1 - \mu z)^3$ we get

$$\hat{g}_4 \;=\; \hat{b}^T \left[\, -\mu^3 I \;+\; 3\mu^2 A \;-\; 3\mu A^2 \;+\; A^3 \,\right] e\,. \qquad (46)$$

Using (34)—(37) and (42)—(44) in (46) we obtain

$$\hat{g}_4 \;=\; -\mu^3 \;+\; 3\mu^2/2 \;-\; \mu/2 \;+\; 1/24\,. \qquad (47)$$

We wanted the b-formula to be stricly proper. Since $\text{degree}(H) = 3$, it follows that we must have $g_3 = 0$. Similarly, to attain a proper $\hat{b}$-formula we must have $\hat{g}_4 = 0$. The roots of (41) and (47) are not compatible. ■

Remarks
Using almost the same arguments as used in the proof of Theorem 1 it can be shown that there does not exist an SDIRK formula that is proper with $p = 4$, $s_i = 2$ and $s_e > 1$ Similarly we can show that there does not exist an SDIRK formula that is strictly proper with $p = 4$, $s_i = 3$ and $s_e > 1$.

In the next two theorems we show that for $s_i = 3$ and $s_e < 2$, we cannot get an embedded SDIRK pair with $(p, \hat{p}) = (2, 3)$ where both formulas are stricly proper.

Theorem 4 *No pair of embedded SDIRK formulas exists with the following properties: (a) the b-formula is of order 2 and is strictly proper, (b) the b-formula is of order 3 and is strictly proper, and (c)* $s = s_i = 3$.

Proof Let all diagonal elements of A_3 of (2) be equal to μ. We will show that the requirements lead to conflicting demands on μ.

Expressions (21)–(25) are valid for the b-formula. Analogous expressions are valid for the $\hat{b}$-formula when b is replaced with $\hat{b}$. For an SDIRK with $s = s_i = 3$, $\det(I - zA) = H(z) = (1 - \mu z)^3$. The coefficient of z^3 in the numerator of $\hat{R}(z)$ is given by (33), which is repeated here:

$$\hat{g}_3 \;=\; h_3 \,+\, \hat{b}^T G_2\, e \;=\; h_3 \,+\, \hat{b}^T \Big[\, h_2 I \,+\, h_1 A \,+\, h_0 A^2 \,\Big]\; e \tag{48}$$

An analogous expression applies for the coefficient of z^3 in the numerator of $R(z)$:

$$g_3 \;=\; h_3 \,+\, b^T G_2\, e \;=\; h_3 \,+\, b^T \Big[\, h_2 I \,+\, h_1 A \,+\, h_0 A^2 \,\Big]\; e \tag{49}$$

Both the estimator and auxiliary formulas satisfy order 2 conditions;

$$b^T e = \hat{b}^T e = 1\;, \tag{50}$$

$$b^T c = b^T A e = \hat{b}^T c = \hat{b}^T A e = 1/2\;. \tag{51}$$

Since we want both R and $\hat{R}$ to be stricly proper we must have $g_3 = \hat{g}_3 = 0$. From this and (48)–(51) it follows that

$$b^T A^2 e = \hat{b}^T A^2 e\;. \tag{52}$$

We can combine (50)–(52) to obtain the following:

$$\Big[e \;\; Ae \;\; A^2 e\Big]^T \;[b - \hat{b}] \;=\; [0 \;\; 0 \;\; 0]^T\;. \tag{53}$$

Let $M \equiv [e \;\; Ae \;\; A^2 e]$. If M is nonsingular, then $b = \hat{b}$ and we will not have an embedded pair. So M must be singular. For M singular, $\det(M) = a_{21}^2 a_{32} = 0$. Both $a_{21} = 0$ and $a_{32} = 0$ result in the following dependence on the columns of M:

$$A^2 e \;=\; 2\mu A e \;-\; \mu^2 e\;. \tag{54}$$

Using (54), (50), (51) and $H(z) = (1 - \mu z)^3$ in (49) we obtain

$$\begin{aligned} g_3 \;&=\; -\mu^3 \,+\, 3\mu^2\, b^T e \,-\, 3\mu\, b^T A e \,+\, 2\mu\, b^T A e \,-\, \mu^2\, b^T e \\ &=\; -\mu\,(\,\mu^2 \,-\, 2\mu \,+\, 1/2\,)\,. \end{aligned} \tag{55}$$

Using order 3 conditions (34)–(37) in (48) we obtain

$$\hat{g}_3 \;=\; -\mu^3 + 3\mu^2 - 3\mu/2 + 1/6\;. \tag{56}$$

We want $g_3 = \hat{g}_3 = 0$, but the roots of (55) and (56) are not compatible. ■

Theorem 5 *No pair of embedded SDIRK formulas exists with the following properties: (a) the b-formula is of order 2 and is strictly proper, (b) the $\hat{b}$-formula is of order 3 and is strictly proper, and (c) $s_i = 3$ and $s_e = 1$.*

Proof The proof is similar to that of Theorem 4, hence we will use much directly from there. We will show that the requirements produce conflicting demands on μ, the diagonal element of A_3 of (2).

The denominator of $R(z)$ and $\hat{R}(z)$ is the same as in Theorem 4, i. e. $H(z) = (1-\mu z)^3$, but now $s = 4$. The expressions for g_3 and $\hat{g}_3$ given in (48) and (49) are valid. The coefficient of z^4 in the numerator of $\hat{R}(z)$ can be written as:

$$\hat{g}_4 \;=\; h_4 \,+\, \hat{b}^T G_3\, e \;=\; \hat{b}^T \Big[\, h_3 I \,+\, h_2 A \,+\, h_1 A^2 \,+\, h_0 A^3 \,\Big]\; e\;. \tag{57}$$

By replacing $\hat{b}$ with b we obtain the corresponding expression for g_4. In Theorem 4 it was shown that requiring $g_3 = \hat{g}_3 = 0$ resulted in $b^T A^2 e = \hat{b}^T A^2 e$, see (52). Using (50)–(52) and requiring $g_4 = \hat{g}_4 = 0$ results in

$$b^T A^3 e = \hat{b}^T A^3 e\;. \tag{58}$$

We need $g_4 = \hat{g}_4 = 0$ for both formulas to be strictly proper. From (50)–(52) and (58) we may write

$$\Big[e \;\; Ae \;\; A^2 e \;\; A^3 e\Big]^T \;[b - \hat{b}] \;=\; [0 \;\; 0 \;\; 0 \;\; 0]^T\;. \tag{59}$$

Let $M \equiv [e \;\; Ae \;\; A^2 e \;\; A^3 e]$. If M is nonsingular then $b = \hat{b}$ and we will not have an embedded pair. So we must have a singular M. For M singular, $\det(M) = a_{32}^2 a_{43}(\mu + a_{21}) = 0$. The cases of $a_{32} = 0$, $a_{43} = 0$ and $\mu = -a_{21}$ all result in the same dependence on the columns of M:

$$A^3 e \;=\; 2\mu A^2 e \,-\, \mu^2 A e\;. \tag{60}$$

Using (60) and the order 3 conditions (34)–(37) in (57) we obtain

$$\hat{g}_4 \;=\; -\mu\,(\,\mu^2 \,-\, \mu \,+\, 1/6\,)\;. \tag{61}$$

The expression in (56) is valid for $\hat{g}_3$. For the $\hat{b}$-formula to be strictly proper we require $\hat{g}_4 = \hat{g}_3 = 0$, but the roots of (61) and (56) are not compatible. ■

The next 3 theorems concern SIRK methods. These theorems are comparable to Theorems 4 and 5, i. e. we show that when $s_i = 3$ and $s_e < 2$ we cannot get an

embedded pair of formulas where both are strictly proper. Burrages [2] considered embedded pairs having the following Butcher table:

$$\begin{array}{c|cccccc} c_1 & a_{11} & a_{12} & \dots & a_{1,s-1} & 0 \\ c_2 & a_{21} & a_{22} & \dots & a_{2,s-1} & 0 \\ \vdots & \vdots & & \vdots & \vdots & \\ c_{s-1} & a_{s-1,1} & a_{s-1,2} & \dots & a_{s-1,s-1} & 0 \\ c_s & a_{s,1} & a_{s,2} & \dots & a_{s,s-1} & \mu \\ \hline & b_1 & b_2 & \dots & b_{s-1} & 0 \\ \hline & \hat{b}_1 & \hat{b}_2 & \dots & \hat{b}_{s-1} & \hat{b}_s \end{array} \tag{62}$$

Let A_{s-1} be the $(s-1)\times(s-1)$ upper left block of the A matrix of (62). This was found from

$$A_{s-1} = V_{s-1}\,\breve{A}_{s-1}\,V_{s-1}^{-1} \tag{63}$$

where V_{s-1} is an $(s-1)\times(s-1)$ Vandermonde matrix defined in terms of the c_i and $\breve{A}_{s-1}$ is an $(s-1)\times(s-1)$ matrix all of whose eigenvalues are equal to μ. Details can be found from Burrage's paper.

Theorem 6 *Consider the embedded pair of SIRK formulas with $s = s_i = 3$ presented in equations (33)–(34) of Burrage's 1978 paper [2]. The pair has one parameter μ. There is no value of μ for which both the b-formula and the $\hat{b}$-formula are strictly proper.*

Proof The stability functions for the two formulas are as follows:

$$R(z) = \frac{2 - 4\,z\mu + 2\,z^2\mu^2 + 2\,z - 4\,z^2\mu + z^2}{2 - 4\,z\mu + 2\,z^2\mu^2}$$

$$\hat{R}(z) = \frac{6\,z^3\mu^3 - 18\,z^2\mu^2 - 18\,z^3\mu^2 + 18\,z\mu + 9\,z^3\mu + 18\,z^2\mu - 6 - 6\,z - z^3 - 3\,z^2}{-6 + 18\,z\mu - 18\,z^2\mu^2 + 6\,z^3\mu^3}$$

For the b-formula to be strictly proper the degree of the numerator in $R(z)$ must be less than the degree of the denominator. From this it follows that:

$$2\mu^2 - 4\mu + 1 = 0\ .$$

A similar condition on $\hat{R}(z)$ results in

$$6\mu^3 - 18\mu^2 + 9\mu - 1 = 0\ .$$

These two polynomials in μ do not have any common roots. ■

One type of embedded formulas not considered by Burrage [2] was that corresponding to the following Butcher table:

$$\begin{array}{c|c} c & V_s\,\breve{A}_s\,V_s^{-1} \\ \hline & b^T \\ \hline & \hat{b}^T \end{array} \tag{64}$$

We do not assume that $b_s = 0$ as in Burrage's pair (62). In the next theorem we consider this embedded pair.

Theorem 7 *No embedded pair of SIRK formulas exists with the following properties: (a) its Butcher table is given (64) with V_s and A_s defined in Burrage's paper [2], (b) $s = s_i = 3$, (c) the b-formula is of order 2 and is strictly proper, and (d) the $\hat{b}$-formula is of order 3 and is strictly proper.*

Proof This theorem is analogous to Theorem 4 and hence we will use some of the material there. As before we will show that the requirements lead to impossible demands.

As in the proof of Theorem 4, we find that for b and $\hat{b}$ to be different M must be singular, where $M \equiv [e \;\; Ae \;\; A^2e]$. For M to be singular we require

$$\det(M) = (1/2)\,(c_3 - c_2)\,(c_1 - c_3)\,(c_1 - c_2) = 0 \tag{65}$$

By using the order 3 conditions, (34)–(37), we can obtain the following expressions for the elements of $\hat{b}$:

$$\hat{b}_1 = \frac{6\,c_2\,c_3 - 3\,c_2 - 3\,c_3 + 2}{6\,(c_1 - c_3)\,(c_1 - c_2)}\;, \quad \hat{b}_2 = -\frac{6\,c_3\,c_1 - 3\,c_1 - 3\,c_3 + 2}{6\,(c_1 - c_2)\,(c_2 - c_3)}\;,$$

$$\hat{b}_3 = \frac{6\,c_2\,c_1 - 3\,c_1 - 3\,c_2 + 2}{6\,(c_2 - c_3)\,(c_1 - c_3)}$$

If we satisfy (65), then at least two of the $\hat{b}_i$ coefficients will be undefined. ■

Our next theorem concerns a SIRK-based embedded pair that has one explicit stage. The Butcher table corresponding to this pair is as follows:

$$\begin{array}{c|cccc} c_1 & a_{11} & a_{12} & a_{1,3} & 0 \\ c_2 & a_{21} & a_{22} & a_{2,3} & 0 \\ c_3 & a_{31} & a_{32} & a_{33} & 0 \\ c_4 & a_{41} & a_{42} & a_{43} & 0 \\ \hline & b_1 & b_2 & b_3 & b_4 \\ \hline & \hat{b}_1 & \hat{b}_2 & \hat{b}_3 & \hat{b}_4 \end{array} \tag{66}$$

In (66) the upper left 3×3 block of the A matrix is given by

$$A_3 = V_3\,\breve{A}_3\,V_3^{-1} \tag{67}$$

where V_3 and $\breve{A}_3$ are taken from Burrage's paper [2]. All three eigenvalues of A_3 are equal to μ.

Theorem 8 *Consider the embedded pair of formulas given in (66) and (67). Let this pair have the following properties: (a) the b-formula is of order 2 and is strictly proper, (b) the $\hat{b}$-formula is of order 3 and is strictly proper. These conditions imply that $R(z) = \hat{R}(z)$.*

Proof Expressions (21)–(25) are valid for the b-formula. Analogous expressions are valid for the $\hat{b}$-formula when b is replaced with $\hat{b}$. Combining (21)–(25) with (66) we get

$$\hat{R}(z) \;=\; \frac{\hat{g}_0 + \hat{g}_1 z + \hat{g}_2 z^2 + \hat{g}_3 z^3 + \hat{g}_4 z^4}{\det(I - zA)} \;=\; \frac{\hat{g}_0 + \hat{g}_1 z + \hat{g}_2 z^2 + \hat{g}_3 z^3 + \hat{g}_4 z^4}{H(z)} \qquad (68)$$

From (66) it follows that degree$(H) \leq 3$. For $\hat{R}(z)$ to be strictly proper we require that $\hat{g}_3 = \hat{g}_4 = 0$. Expressions for $\hat{g}_0$, $\hat{g}_1$ and $\hat{g}_2$ can be found from (27) and (25):

$$\hat{g}_0 = h_0 \;, \quad \hat{g}_1 = h_1 + h_0 \hat{b}e \;, \quad \hat{g}_2 = h_2 + h_1 \hat{b}e + h_0 \hat{b}Ae \;. \qquad (69)$$

Using *only the order 2 conditions*, $\hat{b}e = 1$ and $\hat{b}Ae = 1/2$, we can rewrite (69):

$$\hat{g}_0 = h_0 \;, \quad \hat{g}_1 = h_1 + h_0 \;, \quad \hat{g}_2 = h_2 + h_1 + h_0/2 \;. \qquad (70)$$

The expression for $R(z)$ is completely analogous to (68) when $\hat{g}_j$ is replaced by g_j. We want $R(z)$ to be strictly proper, so we require $g_3 = g_4 = 0$. The expressions derived for $\hat{g}_0$, $\hat{g}_1$ and $\hat{g}_2$ in (70) used only order 2 conditions. It follows that precisely the same expressions would be obtained for g_0, g_1 and g_2. Hence $R(z) = \hat{R}(z)$. ■

Remarks

1. In proving Theorem 8 we never actually used the fact that the embedded pair in question was a SIRK pair, i. e. that all three nonzero eigenvalues of A were equal to μ. Hence the result is true for any IRK having 3 implicit stages and one explicit.

2. A two-parameter family of embedded SIRK formulas that meets the conditions of Theorem 8 is given in Fig. 1. The free parameters remaining are $\breve{a}_{42}$ and b_4. To prevent the b-formula from becoming an order 3 formula we must have $\breve{a}_{42} \neq 1/2$ and $b_4 \neq 0$. The purpose of this SIRK pair is to serve as an example; there is nothing special about it. In particular several free parameters were set in rather arbitrary fashion: (a) c_1, c_2, c_3 and $c_4 = \breve{a}_{41}$ were given values such that the interval $[0, 1]$ was nicely divided, and (b) either $b_4 = 0$, or $\hat{b}_4 = 0$, but not both, and in Fig. 1 we chose the latter.

Theorem 8 can be extended to any embedded IRK pair whose A matrix is given by (2).

Theorem 9 *Consider an embedded pair of IRK formulas whose A matrix is given by (2). Let this pair have the following properties: (a) A_3 is a 3×3 matrix, (b) the b-formula is of order 2 and is strictly proper, (c) the $\hat{b}$-formula is of order 3 and is strictly proper. These conditions imply that $R(z) = \hat{R}(z)$.*

Proof Expressions (21)–(25) are valid for the b-formula. Analogous expressions are valid for the $\hat{b}$-formula when b is replaced with $\hat{b}$. Using the fact that A_1 and A_6 in A of (2) were triangular with zero diagonals, and that A_3 is a 3×3 nonsingular matrix, we can write

$$\det(I - zA) \;=\; \det(I - zA_1)\; \det(I - zA_3)\; \det(I - zA_6) \;=\; \det(I - zA_3) \;=\; H(z) \;.$$

$$V_3 = \begin{bmatrix} 1 & c_1 & {c_1}^2 \\ 1 & c_2 & {c_2}^2 \\ 1 & c_3 & {c_3}^2 \end{bmatrix} \quad A_3 = \begin{bmatrix} 0 & 0 & 2\,\mu^3 \\ 1 & 0 & -6\,\mu^2 \\ 0 & 1/2 & 3\,\mu \end{bmatrix} \quad \begin{bmatrix} a_{41} \\ a_{42} \\ a_{43} \end{bmatrix} = (V_3^{-1})^T \begin{bmatrix} \breve{a}_{41} \\ \breve{a}_{42} \\ \breve{a}_{43} \end{bmatrix}$$

$$\hat{b} = [2/3 \quad -1/3 \quad 2/3 \quad 0]^T \ , \quad c = [1/4 \quad 1/2 \quad 3/4]^T$$

$$\mu = 1 \ - \ \cos(\varpi/3)/\sqrt{2} \ + \ \sqrt{3}\sin(\varpi/3)/\sqrt{2} \quad , \quad \varpi = \arctan(\sqrt{2}/4)$$

$$\breve{a}_{41} = 1 \ , \quad \breve{a}_{43} = 2\mu\,(\,\mu^2 - 3\mu + 3a_{42}\,)$$

$$b = [b_3 + 2b_4\,, \ \ 1 - 2b_3 - 3b_4\,, \ \ b_3\,, \ \ b_4]^T \ ,$$

$$b_3 = 5\,b_4 + 16\,\mu^3 - 16\,b_4\,a42 - 48\,\mu^2 + 24\,\mu - 2$$

Figure 1: An embedded SIRK pair satisfying the conditions of Theorem 8. The Butcher table for this method is given in (66).

Since A_3 is a 3×3 nonsingular matrix, it follows that degree$(H) = 3$. For $R(z)$ and $\hat{R}(z)$ to be strictly proper, the numerator polynomials must obey $g_j = \hat{g}_j = 0, j > 2$. Using exactly the same arguments as in Theorem 8, we may prove that $g_j = \hat{g}_j, j = 0, 1, 2$. It follows that $R(z) = \hat{R}(z)$. ■

Remarks
Consider the local error estimate provided by (8) and (9) for the linear problem (3). If $R(z) = \hat{R}(z)$, then this error estimate is always 0, and hence it is useless. The same conclusion can be drawn if λ in (3) is a matrix. Hence any embedded pairs fulfilling the conditions of Theorem 9 cannot be used on linear problems.

5 Error coefficients

In this section we will make some comments on the truncation error coefficients of implicit Runge-Kutta methods. Suppose we are numerically integrating a set of ODE's of the form

$$\mathbf{y}' = \mathbf{f}(\mathbf{y}, t) \ . \tag{71}$$

For smooth $\mathbf{f}$ the local truncation error (lte) made when we advance the solution from t_n to $t_n + h$ using the estimator formula can be given by

$$\text{lte} = \sum_{i=p+1}^{\infty} h^i \left(\sum_{j=1}^{\sigma_i} T_{ij}\, D_{ij} \right) \ , \tag{72}$$

where p is the order of the method. Similarly, assuming $\hat{p} = p + 1$, the local truncation error for the auxiliary formula is given by

$$\widehat{\text{lte}} = \sum_{i=p+2}^{\infty} h^i \left(\sum_{j=1}^{\sigma_i} \widehat{T}_{ij}\, D_{ij} \right) \ . \tag{73}$$

In (72) and (73) the D_{ij} are elementary differentials or sums and products of partial differentials of f. For more details one should consult Hosea's paper [6]. Our interest however is with the truncation error coefficients, T_{ij} and $\widehat{T}_{ij}$, and the desirable properties they should have.

The usual assumption is that the h^{p+1} term in (72) dominates the other terms. Given that this assumption is true, we would like the coefficients of h^{p+1} in (72), $[\, T_{p+1,1}\,,\, T_{p+1,2}\,,\, \ldots\,, T_{p+1,\sigma_{p+1}}\,]$, to be "small". The smaller these coefficients, the more accurate the results from the estimator formula will be for a given h. Now we address the assumption that the h^{p+1} term in (72) dominates.

Let $\mathbf{y}_{n+1}$ and $\hat{\mathbf{y}}_{n+1}$ be the results provided by the estimator and auxiliary formulas respectively at $t_{n+1} = t_n + h$. The error estimate can then be written as follows

$$\delta_{n+1} \;\equiv\; \hat{\mathbf{y}}_{n+1} - \mathbf{y}_{n+1} \;=\; -h^{p+1} \sum_{j=1}^{\sigma_{p+1}} T_{ij}\, D_{ij} \;+\; \sum_{i=p+2}^{\infty} h^i \left(\sum_{j=1}^{\sigma_i} (\, \widehat{T}_{ij} - T_{ij}\,)\, D_{ij} \right) \;. \tag{74}$$

The error estimate is typically used to control h. One standard controller used to adjust h is given by

$$h_{n+1} = \left(\frac{\varphi}{r_{n+1}} \right)^{1/(p+1)} h_n \;, \tag{75}$$

where r_n is some scalar estimate of (relative) error obtained from δ_n and φ is a desired value for r_n provided by the user. This controller *assumes* the local error estimate behaves like a formula of order $p+1$. For this assumption to be reasonable, and hence for the controller to perform reasonably, we have two objectives: (i) the $i = p+1$ term in the summation of (72) dominates the terms $i > p+1$, and (ii) the first summation on the right hand side of (74) should dominate the second summation. The D_{ij} terms arise from $\mathbf{f}$, which we cannot affect when designing an embedded pair. However we can affect the T_{ij} and $\widehat{T}_{ij}$ terms. Lacking any better knowledge, we will assume that all D_{ij} are the same size. With this assumption, the two objectives above can be expressed in terms of these truncation coefficients as follows:

$$\|\, T_{p+1,\bullet}\,\| \;>\; h^j \|\, T_{j+p,\bullet}\,\|,\; j = 2, 3, \ldots \tag{76}$$

$$\|\, T_{p+1,\bullet}\,\| \;>\; h^j \|\, \widehat{T}_{j+p,\bullet} - T_{j+p,\bullet}\,\|,\; j = 2, 3, \ldots \tag{77}$$

(The notation $T_{i\bullet}$ means row i of matrix T.) When designing an embedded pair we cannot know what values of h are to be used. What we can hope for is that the following ratios,

$$\kappa_1(j) \;\equiv\; \|\, T_{j+p,\bullet}\,\| \,/\, \|\, T_{p+1,\bullet}\,\|\,,\; j = 2, 3, \ldots \tag{78}$$

$$\kappa_2(j) \;\equiv\; \|\, \widehat{T}_{j+p,\bullet} - T_{j+p,\bullet}\,\| \,/\, \|\, T_{p+1,\bullet}\,\|\;,\; j = 2, 3, \ldots \tag{79}$$

are "small". Shampine [11, p. 375] states that in recent work efforts have been made to make (78) small for several j values. Although this is no doubt desirable, in what follows we will restrict ourselves to considering (78) and (79) for $j = 2$. Other workers have also been content with looking at these ratios for only $j = 2$ [7].

Let us consider the consequences when (77) does not hold. We will consider the simple test equation given by (3). For this simple test equation the error estimate

is given by (8) and (9), which when expanded in a Taylor series can be written as

$$\begin{aligned} \delta_{n+1} &= (\,\hat{R}(h\lambda) - R(h\lambda)\,)\,(y_n - q) \\ &= \left[-C_{p+1}(h\lambda)^{p+1} + \sum_{j=p+2}^{\infty} (\hat{C}_j - C_j)(h\lambda)^j \right] (y_n - q)\ . \end{aligned} \tag{80}$$

For this simple test equation the truncation error coefficients corresponding to R and $\hat{R}$ are C_j and $\hat{C}_j$ respectively. For simplicity we will assume that the first term in the summation of (80) is of comparable size to $|C_{p+1}(h\lambda)^{p+1}|$ and all other terms in the summation are much smaller. Using these assumptions we obtain

$$\delta_{n+1} \approx \left[-C_{p+1}(h\lambda)^{p+1} + (\hat{C}_{p+2} - C_{p+2})(h\lambda)^{p+2} \right] (y_n - q)\ . \tag{81}$$

We will use $r_{n+1} = |\delta_{n+1}|/(|y_n| + \eta)$ as our error in the controller (75), where η is some positive scaling factor. Using (81) we may write r_{n+1} as follows:

$$r_{n+1} = \left| (h_n\lambda)^{p+1} \left(-C_{p+1} + h_n\lambda D_{p+2} \right) \right| \zeta_n\ , \tag{82}$$

where $\zeta_n \equiv |y_n - q|/(|y_n| + \eta)$ and $D_{p+2} \equiv (\hat{C}_{p+2} - C_{p+2})$ and the subscript n on h_n indicates that it is the step size used at t_n. Let us suppose that the step at t_n has been accepted, i. e. $r_{n+1} < \varphi$. The next step size given by the controller (75) is

$$h^c_{n+1} = \left(\frac{\varphi}{r_{n+1}} \right)^{p+1} h_n\ .$$

We add a superscript c to emphasize that this is the step size given by the controller. If we use this h^c_{n+1}, what r_{n+2} can we expect? We can find out by substituting h^c_{n+1} into (82):

$$\begin{aligned} r^c_{n+2} &= \left| (h^c_{n+1}\lambda)^{p+1} \left(-C_{p+1} + h^c_{n+1}\lambda D_{p+2} \right) \right| \zeta_n \\ &= \frac{\varphi \left| \, h_n\lambda D_{p+2}(\varphi/r_{n+1})^{1/(p+1)} - C_{p+1} \, \right|}{|\, h_n\lambda D_{p+2} - C_{p+1}\,|} \frac{\zeta_{n+1}}{\zeta_n} \end{aligned} \tag{83}$$

Let us suppose the following hold: $\zeta_{n+1} \geq \zeta_n$, $\mathrm{Re}(\lambda) < 0$, and $C_{p+1}D_{p+2} < 0$. Since we have already assumed that the previous step has been accepted, $r_{n+1} < \varphi$, then from (83) it follows that $r^c_{n+2} > \varphi$. In other words the controller will produce a step size h^c_{n+1} whose corresponding r^c_{n+2} is bigger than the desired φ. Typically, step sizes are accepted if $r < (1 + \tau)\varphi$, for some $\tau > 0$. If we have not chosen τ to be large enough, then the step size will be rejected, which implies wasted computations. The underlying reason for this is that (77) did not hold: i. e. the assumption of the controller that our error estimate was of order $p + 1$ did not hold.

We summarize here what we consider to be desirable properties for the truncation error coefficients:

1. We would like small $T_{p+1,j}$ coefficients, $\forall j$.
2. We would like $\kappa_1(2)$ of (78) to be small.
3. We would like $\kappa_2(2)$ of (78) to be small.

$$A = \begin{bmatrix} \mu\left(4-\sqrt{2}\right)/4 & \mu\left(4-3\sqrt{2}\right)/4 & 0 \\ \mu\left(4+3\sqrt{2}\right)/4 & \mu\left(4+\sqrt{2}\right)/4 & 0 \\ \left(-\mu^2\left(11\sqrt{2}+8\right)-\sqrt{2}\right. & \left(\mu^2\left(11\sqrt{2}-8\right)+\sqrt{2}\right. & \mu \\ \left.\quad +4\,\mu\left(1+2\sqrt{2}\right)\right)/8\,\mu & \left.\quad +4\,\mu\left(1-2\sqrt{2}\right)\right)/8\,\mu & \end{bmatrix}$$

$$b^T = \left[\left(4\,\mu\left(1+\sqrt{2}\right)-\sqrt{2}\right)/8\,\mu\,,\ \left(4\,\mu\left(1-\sqrt{2}\right)+\sqrt{2}\right)/8\,\mu\,,\ 0\right]$$

$$\hat{b}^T = \left[\frac{6\,\mu^2\left(2+\sqrt{2}\right)-3\,\mu\left(3+\sqrt{2}\right)+1}{12\,\mu\left(\mu\left(3\sqrt{2}-2\right)-\sqrt{2}\right)},\ \frac{6\,\mu^2\left(-2+\sqrt{2}\right)+3\,\mu\left(3-\sqrt{2}\right)-1}{12\,\mu\left(\mu\left(3\sqrt{2}+2\right)-\sqrt{2}\right)},\ \frac{6\,\mu^2-6\,\mu+1}{21\,\mu^2-18\,\mu+3}\right]$$

Figure 2: The one-parameter embedded SIRK pair of Burrage [2].

6 Analysis of embedded pairs

In this section we analyze embedded pairs of IRK formulas. Some of these embedded pairs come from the literature and some are our own designs. Most of the pairs have formulas of orders 2 and 3, although there are a few exceptions.

We have restricted ourselves to considering pairs of SIRK or SDIRK formulas where $s = 3$. We will try to explain why. Based on the arguments presented in section 3, we consider the L-stability of $R(z)$ and the properness of $\hat{R}(z)$ to be desirable. But $R(z)$ cannot be L-stable if it is not strictly proper (15). From Table 1 we can see that we cannot have these desirable properties if $s_i \leq s < 3$. By restricting ourselves to $s = 3$, from Table 1 it follows that we cannot obtain a pair where both $R(z)$ and $\hat{R}(z)$ are strictly proper, and so they cannot both be L-stable. As was shown in Figure 1, SIRK pairs exists for $s_i = 3$ and $s_e = 1$ for which both $R(z)$ and $\hat{R}(z)$ are L-stable. However the fact that $R(z) = \hat{R}(z)$ for such pairs was disconcerting and raised questions. Is there use for an embedded pair which will only work on problems that are genuinely nonlinear? How can one be sure that even a genuinely nonlinear problem does not act very much like a linear problem for some time range? Because of these uncertainties, we did not investigate such SIRK pairs.

Table 2 contains the literature embedded pairs that we investigated. The list of pairs in Table 2 is only a representative sample of existing pairs; we did not do an exhaustive survey. Table 3 contains pairs that we designed. We shall discuss these in more detail below. Tables 4 and 5 contain analysis of the pairs presented in Tables 2 and 3 respectively. We shall comment on the analysis results at the end of this section.

An explanation is needed for interpreting the stability properties shown in Tables 4 and 5. The stability property shown is the strongest that applies to the

Table 2: Some 3 stage embedded pairs from the literature.

pair	type	A	b	$\hat{b}$	comments	ref.
1	SDIRK	$\begin{bmatrix} 5/6 & 0 & 0 \\ -61/108 & 5/6 & 0 \\ -23/183 & -33/61 & 5/6 \end{bmatrix}$	$\begin{bmatrix} 25/61 \\ 36/61 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 26/61 \\ 324/671 \\ 1/11 \end{bmatrix}$		[9]
2	SDIRK	$\begin{bmatrix} 0 & 0 & 0 \\ \mu & \mu & 0 \\ \varpi & \varpi & \mu \end{bmatrix}$	$\begin{bmatrix} \varpi \\ \varpi \\ \mu \end{bmatrix}$	$\begin{bmatrix} (1-\varpi)/3 \\ (3\varpi+1)/3 \\ \mu/3 \end{bmatrix}$	$\mu = (2-\sqrt{2})/2$ $\varpi = \sqrt{2}/4$	[8]
3	SDIRK	$\begin{bmatrix} \mu & 0 & 0 \\ -0.403494298165 & \mu & 0 \\ -0.381596758045 & a_{32} & \mu \end{bmatrix}$	$\begin{bmatrix} 1.158945191501 \\ -0.158945191501 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0.661090792671 \\ 0.131307259462 \\ 0.207601947867 \end{bmatrix}$	$\mu = 0.43586652150846$ $a_{32} = 1 - a_{31} - \mu$	[3]
4	SIRK	see Fig. 2	see Fig. 2	see Fig. 2	$\mu = (2-\sqrt{2})/2$	[2]
5	SIRK	see Fig. 2	see Fig. 2	see Fig. 2	$\mu = (2+\sqrt{2})/2$	[2]
6	SIRK	see Fig. 2	see Fig. 2	see Fig. 2	$\mu = \cos(\pi/18)/\sqrt{3} + 1/2$	[2]
7	SDIRK	$\begin{bmatrix} \mu & 0 & 0 \\ (1-\mu)/2 & \mu & 0 \\ (-1+16\mu & (5-20\mu & \mu \\ -6\mu^2)/4 & +6\mu^2)/4 & \end{bmatrix}$	$\begin{bmatrix} -\vartheta\mu \\ \vartheta(2\,\mu-1) \\ 0 \end{bmatrix}$	$\begin{bmatrix} a_{31} \\ a_{32} \\ a_{33} \end{bmatrix}$	$\varpi = \arctan(\sqrt{2}/4)/3$ $\mu = 1 - \cos(\varpi)/\sqrt{2} + \sqrt{3}\sin(\varpi)/\sqrt{2}$ $\vartheta = (\mu-1)^{-1}$	[4]

Table 3: Our own designs for 3 stage embedded pairs.

pair	type	A	b	$\hat{b}$	comments	ref.
8	SDIRK	$\begin{bmatrix} \mu & 0 & 0 \\ 1-2\mu & \mu & 0 \\ 1-2\mu & \mu & \mu \end{bmatrix}$	$\begin{bmatrix} a_{31} \\ a_{32} \\ a_{33} \end{bmatrix}$	$\begin{bmatrix} 1/2 \\ 1/2 \\ 0 \end{bmatrix}$	$\mu = 1/2 + \sqrt{3}/6$	[1, p.1008]
9	SDIRK	$\begin{bmatrix} \mu & 0 & 0 \\ 1-2\mu & \mu & 0 \\ -1/2 & 7/100 & \mu \end{bmatrix}$	$\begin{bmatrix} -\varpi(25+7\sqrt{3}) \\ -\varpi(54+43\sqrt{3}) \\ \varpi(100+50\sqrt{3}) \end{bmatrix}$	$\begin{bmatrix} 1/2 \\ 1/2 \\ 0 \end{bmatrix}$	$\mu = 1/2 + \sqrt{3}/6$ $\varpi = 1/21$	[1, p.1008]
10	SDIRK	$\begin{bmatrix} \mu & 0 & 0 \\ 4/9 & \mu & 0 \\ 183/\varpi & -63/\varpi & \mu \end{bmatrix}$	$\begin{bmatrix} a_{31} \\ a_{32} \\ a_{33} \end{bmatrix}$	$\begin{bmatrix} 23/24 \\ -27/56 \\ 11/21 \end{bmatrix}$	$\mu = 2/5$ $\varpi = 200$	
11	SDIRK	$\begin{bmatrix} \mu & 0 & 0 \\ 1/2-\mu & \mu & 0 \\ 2\mu & 1-4\mu & \mu \end{bmatrix}$	$\begin{bmatrix} \varpi \\ 1-2\varpi \\ \varpi \end{bmatrix}$	$\begin{bmatrix} \vartheta \\ 1-2\vartheta \\ \vartheta \end{bmatrix}$	$\mu = 1/2 + \cos(\pi/18)/\sqrt{3}$ $\vartheta = 1/(6(2\mu-1)^2)$ $\varpi = \dfrac{\mu(2\mu^2-4\mu+1)}{8\mu^2-6\mu+1}$	[1, p.1008]
12	SDIRK	$\begin{bmatrix} \mu & 0 & 0 \\ 1/2-\mu & \mu & 0 \\ 2\mu & 1-4\mu & \mu \end{bmatrix}$	$\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$	$\begin{bmatrix} \vartheta \\ 1-2\vartheta \\ \vartheta \end{bmatrix}$	$\mu = 1/2 + \cos(\pi/18)/\sqrt{3}$ $\vartheta = 1/(6(2\mu-1)^2)$	[1, p.1008]
13	SDIRK	$\begin{bmatrix} \mu & 0 & 0 \\ (1-\mu)/2 & \mu & 0 \\ (-1+16\mu & (5-20\mu & \mu \\ -6\mu^2)/4 & +6\mu^2)/4 & \end{bmatrix}$	$\begin{bmatrix} \vartheta(-16\,\mu-9) \\ \vartheta(32\,\mu-7) \\ 9/25 \end{bmatrix}$	$\begin{bmatrix} a_{31} \\ a_{32} \\ a_{33} \end{bmatrix}$	$\varpi = \arctan(\sqrt{2}/4)/3$ $\mu = 1 - \cos(\varpi)/\sqrt{2}$ $+\sqrt{3}\sin(\varpi)/\sqrt{2}$ $\vartheta = (\,25\,(\mu-1))^{-1}$	[1, p.1012]

Table 4: Analysis results for the embedded pairs of Table 2.

pair	stability property for estimator formula	stability property for auxiliary formula	$\gamma = R(z=\infty)$	$\hat{\gamma} = \hat{R}(z=\infty)$	$T_{3\bullet}$	$T_{4\bullet}$	$\hat{T}_{4\bullet}$	$\kappa_1(2)$	$\kappa_2(2)$
1	S-stable	S-stable	−6.8e−1	−7.3e−1	[3.1e−3, 2.5e−2]	[2.3e−4, −7.4e−2, 2.1e−2, 1.4e−1]	[−1.1e−3, −8.5e−2, 4.3e−3, 8.5e−2]	6.3e0	2.2e0
2	strongly S-stable	not proper	0	∞	[−5.5e−2, −4.5e−2]	[−2.0e−2, −4.4e−2, −2.5e−2, −1.8e−2]	[1.5e−3, −4.2e−2, −2.5e−2, −1.3e−2]	8.0e−1	3.1e−1
3	S-stable	L-stable S-stable	−9.6e−1	0	[5.7e−2, −1.4e−1]	[2.6e−2, −4.9e−2, −3.8e−2, −6.5e−2]	[−2.1e−3, 1.5e−2, −1.6e−2, 4.4e−2]	6.4e−1	8.9e−1
4	strongly S-stable	proper	0	1.6e0	[−4.0e−2, 0]	[−2.5e−2, 0, −1.2e−2, 0]	[4.1e−4, 0, −1.6e−3, 0]	6.8e−1	6.8e−1
5	L-stable, S-stable	proper	0	−2.8e−1	[1.4e0, 0]	[3.3e0, 0, 2.3e0, 0]	[4.7e−1, 0, −1.9e0, 0]	2.9e0	3.7e0
6	S-stable	S-stable	−4.3e−1	−6.3e−1	[2.4e−1, 0]	[3.4e−1, 0, 4.3e−1, 0]	[0, 0, 0, 0]	2.3e0	2.3e0
7	S-stable	strongly S-stable	−9.6e−1	0	[3.5e−2, −1.1e−1]	[1.7e−2, 6.1e−2, −3.9e−2, −5.6e−2]	[−7.9e−3, 3.2e−2, 1.7e−2, 1.7e−2]	7.8e−1	1.1e0

Table 5: Analysis results for the embedded pairs of Table 3.

pair	stability property for estimator formula	stability property for auxiliary formula	$\gamma = R(z=\infty)$	$\hat{\gamma} = \hat{R}(z=\infty)$	$T_{3\bullet}$	$T_{4\bullet}$	$\hat{T}_{4\bullet}$	$\kappa_1(2)$	$\kappa_2(2)$
8	strongly S-stable	S-stable	0	−7.3e−1	[−6.6e−2, 4.2e−1]	[−4.4e−2, 1.9e−1, 1.7e−1, 8.1e−1]	[0, −9.0e−2, 0, 9.0e−2]	2.0e0	1.9e0
9	L-stable, S-stable	S-stable	0	−7.3e−1	[2.8e−1, 7.8e−2]	[1.3e−1, 5.4e−1, 2.7e−1, 5.4e−1]	[0, −9.0e−2, 0, 9.0e−2]	2.8e0	2.9e0
10	strongly S-stable	S-stable	0	−2.7e−1	[5.8e−3, −2.3e−2]	[−3.1e−3, 1.3e−2, 7.3e−3, −7.3e−3]	[−7.5e−3, 3.4e−2, 2.0e−2, 5.0e−3]	7.1e−1	1.2e0
11	L-stable, S-stable	S-stable	0	−6.3e−1	[1.3e−1, 6.4e−1]	[6.7e−2, 3.2e−1, 6.8e−1, 1.7e0]	[0, 0, 0, 0]	2.9e0	2.9e0
12	S-stable	S-stable	−4.3e−1	−6.3e−1	[4.2e−2, 2.0e−1]	[2.1e−2, 9.9e−2, 2.1e−1, 5.4e−1]	[0, 0, 0, 0]	2.9e0	2.9e0
13	S-stable	strongly S-stable	−1.7e−1	0	[6.0e−3, −2.0e−2]	[−3.6e−3, 1.6e−2, 6.9e−3, 4.6e−3]	[−7.9e−3, 3.2e−2, 1.7e−2, 1.7e−2]	8.9e−1	1.1e0

formula. The following implications show the ranking:

$$\text{strongly S}-\text{stable} \;\Rightarrow\; \text{L}-\text{stable} \;\Rightarrow\; \text{A}-\text{stable}\,,\ \text{strictly proper } R(z)$$

$$\text{S}-\text{stable} \;\Rightarrow\; \text{A}-\text{stable} \;\Rightarrow\; \text{proper } R(z)$$

We shall now discuss the pairs that we designed shown in Table 3. We restricted ourselves to SDIRK designs, hence all comments regarding pairs 8–13 apply only to SDIRK pairs. We do not take all the credit (or blame!) for the pairs 8–13. Only pair 10 is completely our own design. For all other pairs we took a single existing SDIRK formula and added another SDIRK formula to the existing one. These existing formulas all had some desirable properties as will become evident. We next give a brief description of how these pairs were designed. In designing these pairs, if we had free parameters left over, we *did not* use any optimization routine to find the final values; we selected the values ourselves. The "truncation coefficients demands" referred to are those presented at the end of section 5..

Pairs 8 and 9 The starting point for these two pairs was the unique 2-stage order 3 A-stable SDIRK formula [1]. This order 3 SDIRK formula corresponds to $\hat{b}$ and the upper right 2×2 block of A in Table 3. For pair 8, we required the estimator formula to be L-stable and stiffly accurate. These requirements left no free parameters. For pair 9 we no longer required the stiff accuracy property, which meant that after satisfying L-stability and order requirements we had two free parameters left. We used these parameters to try to satisfy the truncation coefficient demands. Pairs 8 and 9 are peculiar when compared to "traditional" embedded pairs in the following sense. In traditional embedded pairs, e. g. pairs 1, 3 and 7 Table 2, the order 2 formula uses the first 2 stages and the order 3 formula uses all 3 stages, i. e. $b_3 = 0$. In pairs 8 and 9, the order 3 formula uses the first 2 stages and the order 2 formula uses all 3 stages, i. e. $\hat{b}_3 = 0$. The reason for this is as follows: given an L-stable estimator formula with $b_3 = 0$, it is impossible to find an A-stable 3 stage auxilairy formula. However in pairs 8 and 9 the auxiliary formula is A-stable.

Pair 10 Pair 10 is our completely our own design. After satisfying L-stability and stiff accuracy requirements we had one parameter left. We used this parameter to try to satisfy the demands on the truncation coefficients. Finally, we checked that $\hat{\gamma} = \lim_{|z|\to\infty} |\hat{R}(z)|$ was significantly less than 1.

Pairs 11 and 12 These pairs are based on the unique 3-stage A-stable SDIRK formula of order 4. The rationale behind using an order 4 formula for an auxiliary formula is as follows. Typically we use an order 3 formula and estimate the error of the order 2 formula by essentially assuming that the order 3 formula provides an exact value (see section 5). This assumption should be much better if the auxiliary formula is order 4. In pair 11 we required the estimator formula to be L-stable. After satisfying this requirement there were no free parameters. In pair 12 we did not require the estimator formula to be L-stable. With the one free parameter left we tried to make the truncation coefficients better than those obtained in pair 11. We had to reach some compromise between improving the truncation coefficients demands and not making $\hat{\gamma}$ too close to γ (see section 3).

Pair 13 Pair 13 is based on the unique 3 stage order 3 SDIRK that is strongly S-stable. (So is pair 7.) We use this as the auxiliary formula. From Theorem 4 we know that it is impossible to obtain an L-stable estimator formula to go along with this auxiliary formula. After satisfying the order conditions on the estimator formula, we have one parameter left. We use this one parameter to strike some sort of compromise between the truncation coefficient demands and the size of $\hat{\gamma}$.

We now present some comments on Tables 2–5.

1. Most of the pairs from the literature, Table 2, have $b_3 = 0$, i. e. the estimator formula does not use the third stage. In designing our pairs, Table 3, we have not restricted ourself to pairs with $b_3 = 0$. Our reasoning was that since three stages are needed to get an error estimate, then one may as well use all three stages in both estimator and auxiliary formulas. If we had not allowed b_3 to be nonzero, it would have not been possible to obtain L-stable estimator formulas in pairs 8 and 11.

2. We will give an example of the compromises that had to be made when designing the pairs in Table 3. Our example concerns pairs 11 and 12. By removing the L-stability requirement of pair 11, we gained one free parameter, in our case b_3. With b_3 we tried to improve the truncation error coefficients, in particular we tried to make $\|T_{3\bullet}\|$ smaller than in pair 11. The problem was that improving $\|T_{3\bullet}\|$, also brought γ closer to $\hat{\gamma}$, which as discussed in section 3 is not desirable. The reason for this soon became obvious: $\|T_{3\bullet}\|$ got smaller when b was made closer to $\hat{b}$. We chose $b_3 = 0$ to strike a compromise between having a small $\|T_{3\bullet}\|$ and having γ "far" from $\hat{\gamma}$. If we could allow γ to be closer to $\hat{\gamma}$ than in pair 12, we could get a much better $\|T_{3\bullet}\|$ value. For example, when $b = [1/10\,,\ 4/5\,,\ 1/10]^T$, then $T_{3\bullet} = [9.3\text{e}{-3}\,,\ 4.4\text{e}{-2}]$ but $\gamma = -5.9\text{e}{-1}$ and $\hat{\gamma} = -6.3\text{e}{-1}$.

3. All the SIRK pairs, 4–6 in Table 2, have $T_{32} = 0$. According to Shampine [11, p. 374] this is not a desirable property: if $T_{31} = 0$ or $T_{32} = 0$ '... there would be a class of problems for which the formula is of order three rather than order two. This causes obvious difficulties with error estimation ...'

4. Pairs 6, 11 and 12 have an order 4 auxiliary formula. Hence for these pairs $\hat{T}_{4\bullet}$ is the zero vector.

5. Pair 11 has an L-stable estimator formula and an order 4 auxiliary formula. Burrage [2] was not able to obtain such a pair with his SIRK formulation (see pair 6).

6. When using an RK formula to advance the states from t_n to t_n+h, the function $\mathbf{f}(\mathbf{y}, t)$ is evaluated at $t_n + c_i h$, $i = 1, 2, \ldots\ s$, where c_i is the sum of row i of A. In general it is desirable that $0 \le c_i \le 1, \forall i$. Except for pairs 5, 6, 11 and 12, all pairs have $0 \le c_i \le 1, \forall i$.

7. Hosea and Shampine [8] do realize that $\hat{R}(z)$ of pair 2 is not proper. To compensate for this they propose using one additional "Rosenbrock" stage on the error estimate. One should consult their paper for more details.

8. We have been assuming that the lower order estimator formula was the one used to advance the state, $\mathbf{y}$, since typically the estimator formula had better stability properties. This reasoning does not hold for pairs 3, 7 and 13. Cash [4] advances using the auxiliary formula when using pair 7 and is of the opinion that it is beneficial. There are different opinion on this matter though. One should consult Shampine [11, p. 342] and Thomas and Gladwell [12] for discussion on this.

9. Based on the previous, we will ignore pairs 3, 7 and 12 in this comment. We believe that pair 10 compares quite favorably with all other pairs. Its stability properties are good and its $\|T_{3\bullet}\|$, $\kappa_1(2)$, $\kappa_2(2)$ and $\hat{\gamma}$ values are certainly "small" compared to the averages of these measures for all other pairs. In addition to the extra freedom provided by not assuming $b_3 = 0$ (see comment 1 above), we believe the main reason for the goodness of pair 10 is that the two formulas were designed together, rather than as separate formulas. All other pairs in Table 3 were based on existing SDIRK formulas, that by themselves had desirable properties. But trying to "staple" on another SDIRK formula to create an embedded pair seems to result in something which is comparatively poor. For example in pairs 11 and 12 we cannot attain simulataneously a small $\|T_{3\bullet}\|$ value and an estimator formula which is L-stable. In a sense, some of the literature pairs suffer from this same defect. For example pair 2 results from the following conditions: (a) $a_{11} = 0$, (b) the estimator formula is strongly S-stable, (c) $0 \leq c_k \leq 1, k = 1, 2, 3$, (d) $p = 2$, and (e) $\hat{p} = 3$. Note that the only condition related explicitly to the auxiliary formula is (e). Similarly setting $b_3 = 0$ and requiring strong S-stability of the estimator formula in pair 4 fixes both the estimator and auxiliary formulas. The discussion in sections 3 and 5 were meant to show that merely satisfying order conditions on the auxiliary formula does not necessarily result in a good pair.

References

[1] Alexander, R. Diagonally implicit Runge-Kutta methods for stiff O. D. E. 's. *SIAM J. Numer. Anal.* **14**, 1006-21, 1977.

[2] Burrage, K. A special family of Runge-Kutta methods for solving stiff differential equations. *BIT*, **18**, 22-41, 1978.

[3] Cameron, I.T. Solution of differential-algebraic systems using diagonally implicit Runge-Kutta methods. *IMA J. of Numer. Anal.*, **3**, 273-289, 1983.

[4] Cash, J.R. Diagonally implicit Runge-Kutta formulae with error estimates, *J. Inst. Maths. Applics.*, **24**, 293-301, 1979.

[5] Hairer, E. and Wanner G. *Solving ordinary differential equations, Vol II, Stiff and differential-algebraic problems*, Springer-Verlag, Berlin, 1991.

[6] Hosea, M. E. A new recurrence for computing Runge-Kutta truncation error coefficients. *SIAM J. Numer. Anal.*, **32**, 1995.

[7] Hosea, M. E. and Shampine, L. F. Estimating the error of the classic Runge-Kutta formula, *Appl. Maths. and Comp.* **66**, 1994.

[8] Hosea, M. E. and Shampine, L. F. Analysis and implementation of TR-BDF2. To appear in *Applied Numerical Mathematics*. On the internet at http://www.math.niu.edu/∼mhosea/.

[9] Nørsett, S.P. and Thomsen, P.G. Embedded SDIRK-methods of basic order three. *BIT*, **24**, 634-36, 1984.

[10] Prothero, A. and Robinson, A. On the stability and accuracy of one-step methods for solving stiff systems of ordinary differential equations. *Math. Comp.*, **28**, no. 125, 145–62, 1974.

[11] Shampine, L.F. *Numerical solution of ordinary differential equations.* Chapman and Hall, New York, 1994.

[12] Thomas1, R.M. and Gladwell, I. Variable-order variable-step algorithms for second-order systems. Part 1: The methods. *Int. J. Numer. Methods Eng.*, **28**, 39–53, 1988.

[13] Wiberg, D.M. *Theory and problems of state space and linear systems*, Schaum's Outline Series, McGraw-Hill, New York, 1971.