Japan J. Indust. Appl. Math., 18 (2001), 107–130 Area ⟨2⟩

A Functional Fitting Runge-Kutta Method with Variable Coefficients

Kazufumi OZAWA

Graduate School of Information Science, Tohoku University,
Katahira 2-1-1, Aoba-Ku, Sendai 980-8577, Japan
E-mail: ozawa@dais.is.tohoku.ac.jp

Received June 7, 1999
Revised June 23, 2000

In this paper, we propose a functional fitting s-stage Runge-Kutta method which is based on the exact integration of the set of the linearly independent functions $\varphi_i(t)$, $(i = 1, \ldots, s)$. The method is exact when the solution of the ODE can be expressed as the linear combination of $\varphi_i(t)$, although the method has an error for general ODE. In this work we investigate the order of accuracy of the method for general ODEs, and show that the order of accuracy of the method is at least s, if the functions $\varphi_i(t)$ are sufficiently smooth and the method is non-confluent. Furthermore, it is shown that the attainable order of the method is $2s$, like conventional Runge-Kutta methods. Two- and three-stage methods including embedded one of this type are developed.

Key words: variable coefficients, order of accuracy, functional fitting, collocation Runge-Kutta method, embedded Runge-Kutta method

1. Introduction

A number of numerical methods have been developed for solving the initial value problems of ordinary differential equations. Among the methods, variable coefficient methods, whose coefficients are the functions of the parameters of the problem and the stepsize, are particularly useful for the special situations that some knowledge of the problems is known in advance. The variable coefficient methods are classified into three categories. The first one is for the case that the solution of the problem is periodic and its (approximate) frequency is known a priori. For example, the linear multistep methods by Gautschi [4], Bettis [1] and Vanthournout *et al.* [13], and the Runge-Kutta-Nyström method by Ozawa [9] fall into this category. The coefficients of these methods are the functions of the angular frequency ω of the problem, and the stepsize h. The methods of the second category, which are often called exponential fitting, are designed for the solutions of the Schrödinger equations. These methods are based on the exact integration of the functions $\{1, x, \ldots, x^m, x\exp(\pm vx), \ldots, x^p \exp(\pm vx)\}$. The methods by Coleman [3], Simos [11] and Thomas *et al.* [12] are in this category. Runge-Kutta methods in the first or second categories are not yet found. The third class of the methods is efficient when the (approximate) eigenvalues of the Jacobian of the system are known in advance. Nakashima [8] has developed some class of A-stable explicit Runge-Kutta methods, whose coefficients are the functions of the eigenvalues and the stepsize.

The purpose of this paper is to establish the basic theory for the generalized variable coefficient Runge-Kutta (VCRK) method. The method is based on the exact integration of the linearly independent s functions $\{\varphi_m(t)\}_{m=1}^{s}$, where s is the stage of the method. Two- and three-stage VCRK methods including an embedded method are developed using this theory.

2. Functional Fitting Runge-Kutta Method

Let us consider the VCRK

$$\begin{cases} y_{n+1} = y_n + h\sum_{i=1}^{s} b_i(t_n, h)\, f(t_n + c_i h,\, Y_i), \\ Y_i = y_n + h\sum_{j=1}^{s} a_{i,j}(t_n, h)\; f(t_n + c_j h,\, Y_j), \quad i = 1, 2, \ldots, s, \\ t_n = t_0 + nh, \end{cases} \tag{1}$$

for solving the initial value problem

$$y'(t) = f(t, y), \quad y(0) = y_0, \quad t \in [t_0, T]. \tag{2}$$

Throughout this work, we assume that the abscissae c_i are constant, and that the method is non-confluent, that is $c_i \neq c_j$, if $i \neq j$. Consider the case that the coefficients $a_{i,j}(t, h)$ and $b_i(t, h)$ in (1) are determined so as to satisfy the relations

$$\begin{cases} u_m(t + c_i h) = u_m(t) + h\sum_{j=1}^{s} a_{i,j}(t, h)\; u'_m(t + c_j h), \\ u_m(t + h) = u_m(t) + h\sum_{i=1}^{s} b_i(t, h)\; u'_m(t + c_i h), \end{cases} \quad m = 1, 2, \ldots, s, \tag{3}$$

where we choose

$$u'_m(t) = \varphi_m(t), \quad m = 1, 2, \ldots, s.$$

Here we assume that the functions $\varphi_m(t)$, which are also defined on $[t_0, T]$, are linearly independent, that is the Wronskian matrix $W(t)$ associated with $\{\varphi_m(t)\}_{m=1}^{s}$

$$W(t) = \begin{pmatrix} \varphi_1(t) & \varphi_2(t) & \cdots & \varphi_s(t) \\ \varphi'_1(t) & \varphi'_2(t) & \cdots & \varphi'_s(t) \\ \vdots & \vdots & \cdots & \vdots \\ \varphi_1^{(s-1)}(t) & \varphi_2^{(s-1)}(t) & \cdots & \varphi_s^{(s-1)}(t) \end{pmatrix}$$

is nonsingular for all $t \in [t_0, T]$. Note that if we take $\varphi_m(t) = t^{m-1}$ $(m = 1, 2, \ldots, s)$, which are linearly independent for $t > 0$, then the method reduces to the collocation Runge-Kutta method [2]. Thus the VCRK is a generalization of the collocation Runge-Kutta method.

The next theorem guarantees the existence of a unique solution $(a_{i,j}(t,h), b_i(t,h))$ for all $t \in [t_0, T]$ and small $h > 0$.

THEOREM 1. *If the functions* $\{\varphi_m(t)\}_{m=1}^s$ *are sufficiently smooth, then* $a_{i,j}(t,h)$ *and* $b_i(t,h)$ *determined by* (3) *are unique for small* $h > 0$.

Proof. It is clear from (3) that coefficients $a_{i,j}(t,h)$ and $b_i(t,h)$ are uniquely determined when the matrix

$$U(t,h) = \begin{pmatrix} \varphi_1(t+c_1h) & \varphi_1(t+c_2h) & \cdots & \varphi_1(t+c_sh) \\ \varphi_2(t+c_1h) & \varphi_2(t+c_2h) & \cdots & \varphi_2(t+c_sh) \\ \vdots & \vdots & \cdots & \vdots \\ \varphi_s(t+c_1h) & \varphi_s(t+c_2h) & \cdots & \varphi_s(t+c_sh) \end{pmatrix}$$

is nonsingular. This matrix can be expressed as

$$U(t,h) = W^T(t) \begin{pmatrix} 1 & 1 & \cdots & 1 \\ c_1h & c_2h & \cdots & c_sh \\ \vdots & \vdots & \cdots & \vdots \\ \dfrac{(c_1h)^{s-1}}{(s-1)!} & \dfrac{(c_2h)^{s-1}}{(s-1)!} & \cdots & \dfrac{(c_sh)^{s-1}}{(s-1)!} \end{pmatrix} + O(h^s), \tag{4}$$

since $\varphi_m(t)$ are sufficiently smooth. We can easily find from (4) that U is nonsingular for small but nonzero h, since the W is nonsingular by assumption, and the second matrix in the right-hand side is, of course, nonsingular by the assumption that c_i are different from each other. Thus we have proved this theorem. ■

Hereafter we are only concerned with the case that the functions $\{\varphi_m(t)\}$ are linearly independent and sufficiently smooth, i.e. the case that the coefficients of the VCRK are unique. Although the coefficients of the VCRK, in general, depend not only on h and but also on t, we will not consider the dependency on t, since our main concern in this work is the local behavior of the error of the method when h tends to 0, for fixed t. Moreover we simply denote the coefficients by $a_{i,j}$ and b_i, respectively, unless confusions arise.

3. Order of Accuracy of VCRK

It is clear from the construction of the VCRK that the method always gives the exact solution for any $h > 0$, when the solution of ODE (2) is in the class $\mathrm{span}\{\Phi_1(t), \ldots, \Phi_s(t)\}$, where $\Phi_m(t) = \int \varphi_m(t)\, dt$. However, if this is not the case then the method yields an error, and it is important to be able to evaluate the error. Let us consider the case that the VCRK generated by $\{\varphi_m(t)\}_{m=1}^s$ is not exact, that is the solution of the problem is not in the class $\mathrm{span}\{\Phi_1(t), \ldots, \Phi_s(t)\}$. In this case, if the solution $y(t)$, which is assumed to be sufficiently smooth, and the numerical approximation given by the VCRK satisfy for some integer p

$$y_{n+1} - y(t_n + h) = O(h^{p+1}), \quad y(t_n) = y_n, \quad h \to 0, \tag{5}$$

then we say that the VCRK has *order of accuracy p*. Note that this definition is quite the same as that for constant coefficient methods, except that the error is considered in the situation that the coefficients $a_{i,j}$ and b_i given by (3) vary as functions of h when $h \to 0$.

In order to investigate the order of accuracy, we introduce the quantities given by

$$\begin{aligned} B(q) &= \sum_{i=1}^{s} b_i\, c_i^{q-1} - \frac{1}{q}, \\ C_i(q) &= \sum_{j=1}^{s} a_{i,j}\, c_j^{q-1} - \frac{1}{q}\, c_i^q, \quad i = 1, 2, \ldots, s. \end{aligned} \tag{6}$$

These quantities are related to (3), since if we expand $\varphi_m(t)$ into the power series and substitute this expansion into (3), then we have

$$\sum_{q=1} \frac{1}{(q-1)!}\, B(q)\, h^q\, \varphi_m^{(q-1)}(t) = 0, \tag{7}$$

$$\sum_{q=1} \frac{1}{(q-1)!}\, C_i(q)\, h^q\, \varphi_m^{(q-1)}(t) = 0, \quad i = 1, 2, \ldots, s. \tag{8}$$

Let the power series expansions of $a_{i,j}$ and b_i be

$$\begin{aligned} a_{i,j} &= a_{i,j}^{(0)} + a_{i,j}^{(1)}\, h + a_{i,j}^{(2)}\, h^2 + \cdots, \\ b_i &= b_i^{(0)} + b_i^{(1)}\, h + b_i^{(2)}\, h^2 + \cdots, \end{aligned}$$

then the power series expansions of $B(q)$ and $C_i(q)$ are given by

$$\begin{aligned} B(q) &= B^{(0)}(q) + B^{(1)}(q)\, h + B^{(2)}(q)\, h^2 + \cdots, \\ C_i(q) &= C_i^{(0)}(q) + C_i^{(1)}(q)\, h + C_i^{(2)}(q)\, h^2 + \cdots, \end{aligned} \tag{9}$$

where

$$\begin{cases} B^{(l)}(q) = \sum_{i=1}^{s} b_i^{(l)} c_i^{q-1} - \dfrac{1}{q}\delta_{l,0}, \\ C_i^{(l)}(q) = \sum_{j=1}^{s} a_{i,j}^{(l)} c_j^{q-1} - \dfrac{c_i^q}{q}\delta_{l,0}, \end{cases} \qquad l = 0, 1, \ldots,$$

and $\delta_{i,j}$ is the Kronecker delta.

LEMMA 1. *The quantities $B(q)$ and $C_i(q)$ satisfy*

$$B(q) = O(h^{s+1-q}), \quad q = 1, 2, \ldots, s, \tag{10}$$

$$C_i(q) = O(h^{s+1-q}), \quad q = 1, 2, \ldots, s, \quad i = 1, 2, \ldots, s. \tag{11}$$

Proof. We prove only for $B(q)$ since the proof for $C_i(q)$ is done in straightforward manner, if (10) is proved. Let us introduce the following quantity $\beta_{q,l}$ (q, $l = 1, 2, \ldots, s$):

$$\beta_{q,l} = \begin{cases} \dfrac{1}{(q-1)!} B^{(l-q)}(q), & 1 \le q \le l, \\ 0, & l < q. \end{cases} \tag{12}$$

If we substitute (9) into (7) then we have

$$\sum_{l=1}\left(\sum_{q=1}^{l} \varphi_m^{(q-1)}(t)\,\beta_{q,l}\right) h^l = 0. \tag{13}$$

Therefore, the condition that the coefficients of h^l ($l = 1, 2, \ldots, s$) in (13) are being 0 can be written as

$$W^T(t)\,\boldsymbol{B} = 0, \tag{14}$$

where $\boldsymbol{B} = (\beta_{q,l})$. Thus we have $\boldsymbol{B} = 0$. The fact that the qth row of $\boldsymbol{B}$ is equal to 0 leads to

$$B^{(0)}(q) = B^{(1)}(q) = \cdots = B^{(s-q)}(q) = 0, \quad q = 1, 2, \ldots, s,$$

which gives the conclusion. ■

We cannot determine all $b_i^{(l)}$ ($l \ge 0$) completely from (14), since this condition is only a part of (13) that is equivalent to (3), which defines b_i ($i = 1, 2, \ldots, s$) uniquely. However, we can determine only $a_{i,j}^{(0)}$ and $b_i^{(0)}$ from $B^{(0)}(q) = C_i^{(0)}(q) = 0$ ($q = 1, 2, \ldots, s$).

COROLLARY 1. *Coefficients* $a_{i,j}^{(0)}$ *and* $b_i^{(0)}$ *satisfy the so-called simplifying assumption* [2]:

$$\begin{cases} \displaystyle\sum_{j=1}^{s} a_{i,j}^{(0)}\, c_j^{q-1} = \frac{c_i^q}{q}, & i = 1,2,\ldots,s, \quad q = 1,2,\ldots,s, \\ \displaystyle\sum_{i=1}^{s} b_i^{(0)}\, c_i^{q-1} = \frac{1}{q}, & q = 1,2,\ldots,s. \end{cases} \tag{15}$$

It should be noted that although $a_{i,j}^{(0)}$ and $b_i^{(0)}$ are determined by (15) uniquely, and these are just the coefficients of the s-stage collocation Runge-Kutta method defined by c_i [2], these coefficients are independent of the choice of $\{\varphi_m(t)\}$.

LEMMA 2. *Let* $g(t)$ *be an* $(s+1)$*-times continuously differentiable function on* $[t_0, T]$, *then*

$$\begin{aligned} g(t+c_ih) &= g(t) + h\sum_{j=1}^{s} a_{i,j}\, g'(t+c_jh) + O(h^{s+1}) \\ &= g(t) + h\sum_{j=1}^{s} a_{i,j}^{(0)}\, g'(t+c_jh) + O(h^{s+1}), \quad i = 1,2,\ldots,s+1, \end{aligned}$$

where we set $a_{s+1,j} = b_j$ *and* $c_{s+1} = 1$.

Proof. From the Taylor series expansion of $g(t+c_ih)$ we have

$$g(t+c_ih) = g(t) + h\sum_{q=1}^{s} \frac{1}{(q-1)!}\left(\frac{c_i^q}{q}\right) h^{q-1} g^{(q)}(t) + O(h^{s+1}). \tag{16}$$

If we substitute

$$\frac{c_i^q}{q} = \sum_{j=1}^{s} a_{i,j}^{(0)}\, c_j^{q-1} = \sum_{j=1}^{s} a_{i,j}\, c_j^{q-1} + O(h^{s+1-q}), \quad q = 1,2,\ldots,s,$$

into (16) for $i = 1,2,\ldots,s+1$, then we have immediately the conclusion. ■

Next we consider the stage order of the VCRK. The stage order of the VCRK is defined to be the minimal integer $\mu = \min_{i=1}^{s}\{\mu_i\}$, where

$$e_i = Y_i - y(t_n + c_ih) = O(h^{\mu_i+1}), \quad h \to 0,$$

and the condition $y(t_n) = y_n$ is of course assumed.

THEOREM 2. *The stage order of the* s*-stage VCRK is* s.

Proof. If the solution $y(t)$ of (2) is $(s+1)$-times continuously differentiable on $[t_0, T]$, then from Lemma 2 we have

$$y(t_n + c_ih) = y(t_n) + h\sum_{j=1}^{s} a_{i,j}\, y'(t_n + c_jh) + O(h^{s+1}),$$

and therefore the stage errors $e_i = Y_i - y(t_n + c_i h)$ are given by

$$e_i = (1 - a_{i,i}\, h\, f_y)^{-1} h \sum_{j \neq i}^{s} a_{i,j}(e_j\, f_y + O(e_j^2)) + O(h^{s+1}),$$
$$i = 1, 2, \ldots, s, \qquad (17)$$

where f_y is the derivative of $f(t, y)$ with respect to y. It is clear from (17) that $\mu \leq s$. If $\mu < s$ then we have from (17)

$$\mu_i + 1 = \min\Big\{\min_{j \neq i}\{\mu_j\} + 2, s + 1\Big\} \geq \min\{\mu + 2, s + 1\} = \mu + 2,$$

which is a contradiction, so that we have $\mu = s$. ■

THEOREM 3. *The order of accuracy of the VCRK is at least s.*

Proof. Let E denote the local error at $t_{n+1} = t_0 + (n + 1)h$. Then from Lemma 2 we have

$$E = h \sum_{i=1}^{s} b_i \{f(t + c_i h, Y_i) - f(t + c_i h, y(c_i h))\} + O(h^{s+1})$$
$$= h \sum_{i=1}^{s} (b_i\, e_i\, f_y + O(e_i^2)) + O(h^{s+1}),$$

which shows that the order of E is at least $s + 1$. ■

Like the collocation Runge-Kutta method, it has been shown that the stage order of the VCRK is s and the overall order is at least s. For the collocation Runge-Kutta method, if $\sum_{i=1}^{s} b_i c_i^{q-1} = 1/q$ holds even for $q > s$, then the method has higher order (see [2] and [5]). We have already established in Corollary 1 that $\sum_{i=1}^{s} b_i^{(0)} c_i^{q-1} = 1/q$ holds for $1 \leq q \leq s$. If this holds even for $q > s$, it may be possible to obtain higher order VCRK formulae. In the next section we will investigate the higher order formulae.

4. Higher Order VCRK

As we have already seen in the last section that the coefficients $a_{i,j}^{(0)}$ and $b_i^{(0)}$ are uniquely determined by (15), whenever the abscissae c_i are different from each other. Furthermore, if for a fixed integer ν satisfying $1 \leq \nu \leq s$, the relation

$$\int_0^1 t^{q-1} \prod_{i=1}^{s} (t - c_i)\, \mathrm{d}t = 0, \quad q = 1, 2, \ldots, \nu \qquad (18)$$

holds, then $b_i^{(0)}$ satisfy the stronger statement than (15), i.e.

$$\sum_{i=1}^{s} b_i^{(0)} c_i^{q-1} = \frac{1}{q}, \quad q = 1, 2, \ldots, s + \nu, \qquad (19)$$

or equivalently,

$$B^{(0)}(q) = 0, \quad q = 1, 2, \ldots, s + \nu. \tag{20}$$

In this case, the relation

$$\sum_{i=1}^{s} b_i^{(0)} c_i^{\xi-1} a_{i,j}^{(0)} = \frac{b_j^{(0)}}{\xi}(1 - c_j^{\xi}), \quad \xi = 1, 2, \ldots, \nu, \quad j = 1, 2, \ldots, s. \tag{21}$$

is also valid (see IV.5, Lemma 5.4 in [5]), and then the Runge-Kutta method defined by $(a_{i,j}^{(0)}, b_i^{(0)}, c_i)$ is of order $s + \nu$ (see [2], [5] and [7]). For the order of accuracy of the s-stage VCRK, we have the following theorem and corollary:

THEOREM 4. *If the abscissae c_i are taken to satisfy* (18), *then the order of accuracy of the s-stage VCRK is $s + \nu$.*

This theorem shows that the s-stage VCRK always has the same order of accuracy as that of the collocation Runge-Kutta method with the abscissae c_i. Thus we have:

COROLLARY 2. *The attainable order of accuracy of the s-stage VCRK is $2s$.*

Before proving this theorem, we must prove several lemmas. Here we introduce the quantity $D(q, \xi)$ by

$$D(q, \xi) = \sum_{i=1}^{s} b_i^{(0)} c_i^{\xi-1} C_i(q), \quad q = 1, 2, \ldots, s + \nu - \xi, \quad \xi = 1, 2, \ldots, \nu, \tag{22}$$

and investigate the orders of $B(q)$ and $D(q, \xi)$, when the condition (18) holds.

Hereafter we consider $\beta_{q,l}$ not for q, $l = 1, 2, \ldots, s$ but for q, $l = 1, \ldots, s + \nu$, keeping the definition in (12) unchanged. For $\beta_{q,l}$ we have the following lemma:

LEMMA 3. *If for a fixed d $(\nu > d > 0)$,*

$$\beta_{q,\,q+d} = 0, \quad q = 1, 2, \ldots, s, \tag{23}$$

then we have

$$\beta_{q,\,q+d} = 0, \quad q = 1, 2, \ldots, s + \nu - d.$$

Proof. Condition (23) implies

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \\ c_1 & c_2 & \cdots & c_s \\ \vdots & \vdots & & \vdots \\ c_1^{s-1} & c_2^{s-1} & \cdots & c_s^{s-1} \end{pmatrix} \begin{pmatrix} b_1^{(d)} \\ b_2^{(d)} \\ \vdots \\ b_s^{(d)} \end{pmatrix} = 0,$$

which leads to $b_i^{(d)} = 0$ $(i = 1, 2, \ldots, s)$, since c_i $(i = 1, 2, \ldots, s)$ are different from each other. Thus $\beta_{q,q+d} = 0$ for all q satisfying $1 \leq q < q + d \leq s + \nu$. ■

LEMMA 4. *If* (20) *holds, then*

$$\beta_{q,l} = 0, \quad q, l = 1, 2, \ldots, s + \nu.$$

Proof. Let us define the matrix $\boldsymbol{B} = (\beta_{q,l})_{q,l=1}^{s+\nu}$, and divide it into 4 submatrices by

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{B}_{1,1} & \boldsymbol{B}_{1,2} \\ 0 & \boldsymbol{B}_{2,2} \end{pmatrix}, \quad \boldsymbol{B}_{1,1} \in \mathbf{R}^{s\times s}, \quad \boldsymbol{B}_{1,2} \in \mathbf{R}^{s\times\nu}, \quad \boldsymbol{B}_{2,2} \in \mathbf{R}^{\nu\times\nu}.$$

Note that $\boldsymbol{B}_{2,2}$ is strictly upper triangular matrix because of condition (19). For the proof, it suffices to show $\boldsymbol{B}_{1,2} = \boldsymbol{B}_{2,2} = 0$, since we have already had $\boldsymbol{B}_{1,1} = 0$ in Lemma 1. We prove this by induction on the columns of $\boldsymbol{B}_{1,2}$ and $\boldsymbol{B}_{2,2}$.

The condition that the coefficients of h^l $(l = 1, 2, \ldots, s + \nu)$ in (13) is being 0 can be expressed as

$$U\boldsymbol{B} = (U_1\boldsymbol{B}_{1,1},\ U_1\boldsymbol{B}_{1,2} + U_2\boldsymbol{B}_{2,2}) = (0,\ U_1\boldsymbol{B}_{1,2} + U_2\boldsymbol{B}_{2,2}) = 0,$$

where

$$U = \begin{pmatrix} \varphi_1 & \cdots & \varphi_1^{(s-1)} & \varphi_1^{(s)} & \cdots & \varphi_1^{(s+\nu-1)} \\ \varphi_2 & \cdots & \varphi_2^{(s-1)} & \varphi_2^{(s)} & \cdots & \varphi_2^{(s+\nu-1)} \\ \vdots & & \vdots & \vdots & & \vdots \\ \varphi_s & \cdots & \varphi_s^{(s-1)} & \varphi_s^{(s)} & \cdots & \varphi_s^{(s+\nu-1)} \end{pmatrix} = (U_1, U_2),$$

$$U_1 \in \mathbf{R}^{s\times s}, \quad U_2 \in \mathbf{R}^{s\times\nu}.$$

If we denote the two submatrices $\boldsymbol{B}_{1,2}$ and $\boldsymbol{B}_{2,2}$ by their columns, i.e.,

$$\boldsymbol{B}_{1,2} = (\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_\nu), \quad \boldsymbol{B}_{2,2} = (\mathbf{b}_1', \mathbf{b}_2', \ldots, \mathbf{b}_\nu'), \quad \mathbf{b}_i \in \mathbf{R}^{s\times 1}, \ \mathbf{b}_i' \in \mathbf{R}^{\nu\times 1},$$

then the $(s + 1)$st column of $U\boldsymbol{B}$ is

$$(s + 1)\text{st column of } U\boldsymbol{B} = U_1\mathbf{b}_1 + U_2\mathbf{b}_1' = U_1\mathbf{b}_1 = 0,$$

since $\boldsymbol{B}_{2,2}$ is strictly upper triangular and therefore $\mathbf{b}_1' = 0$. Thus we have $\mathbf{b}_1 = 0$ since $U_1 = W^T$ is nonsingular. We have now established that the first columns of $\boldsymbol{B}_{1,2}$ and $\boldsymbol{B}_{2,2}$ are being 0; this completes the proof when $\nu = 1$. Next we show that the second columns are 0 when $\nu > 1$.

From $\boldsymbol{B}_{1,1} = 0$ and $\mathbf{b}_1 = 0$, we have

$$\beta_{1,2} = \beta_{2,3} = \cdots = \beta_{s,s+1} = 0,$$

and therefore we have from Lemma 3

$$\beta_{l,l+1} = 0, \quad l = s+1, \ldots, s+\nu-1.$$

This leads to $\mathbf{b}_2' = 0$ since $\boldsymbol{B}_{2,2}$ is strictly upper triangular. Using this result we have for the $(s+2)$ nd column of $U\boldsymbol{B}$

$$(s+2)\,\text{nd column of } U\boldsymbol{B} = U_1\mathbf{b}_2 + U_2\mathbf{b}_2' = U_1\mathbf{b}_2 = \mathbf{0},$$

which leads to $\mathbf{b}_2 = 0$. Thus we have proved that the second columns of two matrices $\boldsymbol{B}_{1,2}$ and $\boldsymbol{B}_{2,2}$ are 0.

Next we assume that the columns of $\boldsymbol{B}_{1,2}$ and $\boldsymbol{B}_{2,2}$ are all being 0 up to the l th, i.e.,

$$\beta_{i,s+j} = 0, \quad i = 1, 2, \ldots, s+\nu, \quad j = 1, 2, \ldots, l.$$

Since this condition includes

$$\beta_{i,i+d} = 0, \quad i = 1, 2, \ldots, s, \quad d = 1, 2, \ldots, l,$$

l super-diagonals of $\boldsymbol{B}$, i.e. $\beta_{i,j}$ $(0 < j - i \leq l)$, are all 0 because of Lemma 3 and, as a result, for the $(l+1)$ st column of $\boldsymbol{B}_{2,1}$ we have

$$\beta_{s+1,s+l+1} = \beta_{s+2,s+l+1} = \cdots = \beta_{s+l,s+l+1} = 0,$$

that is $\mathbf{b}_{l+1}' = 0$. Therefore we have

$$(l+1)\,\text{st column of } U\boldsymbol{B} = U_1\mathbf{b}_{l+1} + U_2\mathbf{b}_{l+1}' = U_1\mathbf{b}_{l+1} = \mathbf{0},$$

which implies $\mathbf{b}_{l+1} = 0$. Thus we have proved $\boldsymbol{B}_{1,2} = \boldsymbol{B}_{2,2} = 0$. ■

LEMMA 5. *If* $\beta_{q,l} = 0$ *for* $q, l = 1, 2, \ldots, s+\nu$, *then*

$$B(q) = O(h^{\mu_q}), \quad q = 1, 2, \ldots, s+\nu,$$

where

$$\mu_q = \max\{s+\nu+1-q,\ \nu+1\}.$$

Proof. Since ν super-diagonals of $\boldsymbol{B}$, which are of length at least s, are shown to be 0 in the last lemma, we have from Lemma 3 that $b_i^{(d)} = 0$ $(d = 1, 2, \ldots, \nu)$. Therefore $B(q)$ is of order at least $\nu+1$ for any $1 \leq q \leq s+\nu$. On the other hand, from the fact that the qth row of $\boldsymbol{B}$ is equal to 0, we have $B^{(l)}(q) = 0$, $(l = 0, 1, \ldots, s+\nu-q)$, which means that the order of $B(q)$ is $s+\nu-q+1$. Thus we have proved this lemma. ■

COROLLARY 3. *Let $g(t)$ be an $(s+\nu+1)$ times continuously differentiable function on $[t_0, T]$, then the following relation holds:*

$$\begin{aligned} g(t+h) &= g(t) + h\sum_{i=1}^{s} b_i\, g'(t+c_i h) + O(h^{s+\nu+1}) \\ &= g(t) + h\sum_{i=1}^{s} b_i^{(0)}\, g'(t+c_i h) + O(h^{s+\nu+1}). \end{aligned} \tag{24}$$

Let the Taylor series expansion of $D(q,\xi)$, which is defined by (22), be

$$D(q,\xi) = D^{(0)}(q,\xi) + D^{(1)}(q,\xi)\, h + D^{(2)}(q,\xi)\, h^2 + \cdots$$

Then from (15) and (21) we have

$$D^{(0)}(q,\xi) = \sum_{i=1}^{s} b_i^{(0)} c_i^{\xi-1}\left(\sum_{j=1}^{s} a_{i,j}^{(0)}\, c_j^{q-1} - \frac{c_i^q}{q}\right) = 0, \qquad q = 1, 2, \ldots, s+\nu-\xi. \tag{25}$$

LEMMA 6. *If (25) holds then $D(q,\xi)$ satisfies*

$$D(q,\xi) = O(h^{\lambda_{q,\xi}}), \quad \xi = 1, 2, \ldots, \nu, \quad q = 1, 2, \ldots, s+\nu-\xi,$$

where

$$\lambda_{q,\xi} = \max\{s+\nu-\xi+1-q,\ \nu-\xi+1\}.$$

Proof. Here we define the quantity $d_{q,l}$ by

$$d_{q,l} = \begin{cases} \dfrac{1}{(q-1)!}\, D^{(l-q)}(q), & 1 \le q \le l, \\ 0, & l < q. \end{cases} \tag{26}$$

Using the same technique as is used in Lemma 4, we can show $d_{q,l} = 0$ $(q, l = 1, 2, \ldots, s+\nu-\xi)$, and using this result, we can easily prove this lemma. ■

LEMMA 7. *Let $g(t)$ be an $(s+\nu-\xi)$ times continuously differentiable function on $[t_0, T]$, then*

$$\sum_{i,j=1}^{s} b_i^{(0)} c_i^{\xi-1}(a_{i,j} - a_{i,j}^{(0)})\, g(t+c_j h) = O(h^{s+\nu-\xi}), \quad \xi = 1, 2, \ldots, \nu.$$

Proof. Using the Taylor series expansion of $g(t)$, we have

$$\begin{aligned}
&\sum_{i,j=1}^{s} b_i^{(0)} c_i^{\xi-1}(a_{i,j}-a_{i,j}^{(0)})\, g(t+c_j h) \\
&\quad = \sum_{i,j=1}^{s} b_i^{(0)} c_i^{\xi-1}(a_{i,j}-a_{i,j}^{(0)}) \sum_{q=1} \frac{(c_j h)^{q-1}}{(q-1)!}\, g^{(q-1)}(t) \\
&\quad = \sum_{q=1}^{s+\nu-\xi} \frac{h^{q-1}}{(q-1)!}\, g^{(q-1)}(t) \sum_{i=1}^{s} b_i^{(0)} c_i^{\xi-1} C_i(q) + O(h^{s+\nu-\xi}) \\
&\quad = \sum_{q=1}^{s+\nu-\xi} \frac{h^{q-1}}{(q-1)!}\, D(q,\xi)\, g^{(q-1)}(t) + O(h^{s+\nu-\xi}).
\end{aligned}$$

Since in this expression,

$$q-1+\lambda_{q,\xi} = \max\{s+\nu-\xi,\ \nu-\xi+q\} \geq s+\nu-\xi,$$

then we have the conclusion. ■

LEMMA 8. *The stage errors e_i of the VCRK satisfy*

$$\sum_{i=1}^{s} b_i^{(0)} c_i^{\xi-1} e_i = \frac{1}{\xi} \sum_{j=1}^{s} b_j^{(0)} (1-c_j^{\xi})\, e_j\, h\, f_y + O(h^{s+\nu-\xi+1}), \quad \xi = 1,2,\ldots,\nu.$$

Proof. We first evaluate $\sum_{i=1}^{s} b_i^{(0)} c_i^{\xi-1} y(t_n + c_i h)$ for $\xi = 1,2,\ldots,\nu-1$. Assuming $y(t_n) = y_n$, we have from (19) and (21)

$$\begin{aligned}
&\sum_{i=1}^{s} b_i^{(0)}\, c_i^{\xi-1}\, y(t_n + c_i h) \\
&\quad = \sum_{i=1}^{s} b_i^{(0)}\, c_i^{\xi-1} \left(y(t_n) + h \sum_{j=1}^{s} a_{i,j}^{(0)}\, y'(t_n + c_j h) \right) \\
&\qquad + \sum_{i=1}^{s} b_i^{(0)}\, c_i^{\xi-1} \sum_{q=s+1}^{s+\nu-\xi} \frac{h^q\, y^{(q)}(t_n)}{(q-1)!} \left(\frac{c_i^q}{q} - \sum_{j=1}^{s} a_{i,j}^{(0)}\, c_j^{q-1} \right) + O(h^{s+\nu-\xi+1}) \\
&\quad = \frac{1}{\xi}\, y(t_n) + h \sum_{i,j=1}^{s} b_i^{(0)}\, c_i^{\xi-1} a_{i,j}^{(0)}\, y'(t_n + c_j h) \\
&\qquad + \sum_{q=s+1}^{s+\nu-\xi} \frac{h^q\, y^{(q)}(t_n)}{(q-1)!} \left\{ \frac{1}{q} \sum_{i=1}^{s} b_i^{(0)} c_i^{\xi+q-1} - \sum_{i,j=1}^{s} b_i^{(0)}\, c_i^{\xi-1}\, a_{i,j}^{(0)}\, c_j^{q-1} \right\} \\
&\qquad\qquad + O(h^{s+\nu-\xi+1}) \\
&\quad = \frac{1}{\xi}\, y(t_n) + h \sum_{i,j=1}^{s} b_i^{(0)}\, c_i^{\xi-1}\, a_{i,j}^{(0)}\, y'(t_n + c_j h) + O(h^{s+\nu-\xi+1}).
\end{aligned}$$

On the other hand, we have

$$\sum_{i=1}^{s} b_i^{(0)} c_i^{\xi-1} Y_i = \sum_{i=1}^{s} b_i^{(0)} c_i^{\xi-1} \left\{ y(t_n) + h \sum_{j=1}^{s} a_{i,j} f(t_n + c_j h, Y_j) \right\}$$
$$= \frac{1}{\xi} y(t_n) + h \sum_{i,j=1}^{s} b_i^{(0)} c_i^{\xi-1} a_{i,j} f(t_n + c_j h, Y_j).$$

Therefore

$$\sum_{i=1}^{s} b_i^{(0)} c_i^{\xi-1} e_i = h \sum_{i,j=1}^{s} b_i^{(0)} c_i^{\xi-1} (a_{i,j} - a_{i,j}^{(0)}) y'(t_n + c_j h)$$
$$+ h \sum_{i,j=1}^{s} b_i^{(0)} c_i^{\xi-1} a_{i,j}^{(0)} e_j f_y + O(h^{s+\nu-\xi+1}).$$

Applying the result of Lemma 7 to the first term, and (21) to the second term, we have the conclusion. ■

Proof of Theorem 4. Let E be the local error at $t_{n+1} = t_0 + (n+1)h$. Then from the results of Corollary 3 and Theorem 2, we have

$$E = y_{n+1} - y(t_{n+1}) = y_n + h \sum_{i=1}^{s} b_i f(t_n + c_i h, Y_i) - y(t_{n+1})$$
$$= y_n + h \sum_{i=1}^{s} b_i^{(0)} f(t_n + c_i h, Y_i) - y(t_{n+1}) + O(h^{s+\nu+1})$$
$$= h \sum_{i=1}^{s} \left(b_i^{(0)} e_i f_y + O(e_i^2) \right) + O(h^{s+\nu+1})$$
$$= h f_y \sum_{i=1}^{s} b_i^{(0)} e_i + O(h^{s+\nu+1}).$$

Applying the result of Lemma 8 repeatedly to the sum $\sum_{i=1}^{s} b_i^{(0)} c_i^{\xi-1} e_i$ for $\xi = 1, 2, \ldots, \nu$, we have

$$\sum_{i=1}^{s} b_i^{(0)} e_i = (h f_y) \sum_{i=1}^{s} b_i^{(0)} (1 - c_i) e_i + O(h^{s+\nu})$$
$$= (h f_y)^2 \sum_{i=1}^{s} b_i^{(0)} \left(\frac{1}{2} - c_i + \frac{1}{2} c_i^2 \right) e_i + O(h^{s+\nu})$$
$$= \cdots$$

$$= (h f_y)^\nu \sum_{i=1}^{s} b_i^{(0)} Q_\nu(c_i)\, e_i + O(h^{s+\nu}),$$

where $Q_\nu(c_i)$ is a polynomial in c_i of degree ν. Since $e_i = O(h^{s+1})$ we have

$$\sum_{i=1}^{s} b_i^{(0)} e_i = O(h^{s+\nu}),$$

and therefore we have $E = O(h^{s+\nu+1})$. ■

5. Numerical Example

5.1. Two-stage VCRK method

Consider the two-stage VCRK method with $c_1 = \frac{3-\sqrt{3}}{6}$ and $c_2 = \frac{3+\sqrt{3}}{6}$, the zeros of the shifted Legendre polynomial of degree 2. With the choice of the abscissae we have

$$\int_0^1 t^{q-1}(t-c_1)(t-c_2)\,\mathrm{d}t = 0, \quad q = 1, 2, \tag{27}$$

so that $\nu = 2$ in (18) and therefore the VCRK is of order 4, provided that the functions $\varphi_m(t)$ are sufficiently smooth and linearly independent. As the functions $\varphi_m(t)$ we take here

$$\varphi_1(t) = \exp(\mu t)\cos\omega t, \quad \varphi_2(t) = \exp(\mu t)\sin\omega t, \tag{28}$$

where μ and ω are real numbers. These functions are linearly independent for all t if and only if $\omega \neq 0$. In this case the coefficients derived from the functions are independent of t. Here we show the closed formulae of the coefficients together with their power series expansions in h; in order to avoid the cancellations which might occur when h is small the use of the power series expansions is in general preferable.

$$
\begin{cases}
a_{1,1} = \dfrac{\cos\left(\frac{\omega h}{\sqrt{3}}-\theta\right) - e^{-\frac{3-\sqrt{3}}{6}\mu h}\cos\left(\frac{3+\sqrt{3}}{6}\omega h-\theta\right)}{h\sqrt{\mu^2+\omega^2}\sin\left(\frac{\omega h}{\sqrt{3}}\right)} \\
\quad = \dfrac{1}{4} - \dfrac{\mu h}{36} + \dfrac{(\mu^2-3\omega^2)h^2}{288\sqrt{3}} + \dfrac{\mu\left((9-6\sqrt{3})\mu^2+(1+6\sqrt{3})\omega^2\right)h^3}{12960} \\
\quad + \dfrac{(\mu^2+\omega^2)\left((-27+16\sqrt{3})\mu^2+3(3-4\sqrt{3})\omega^2\right)h^4}{155520} + O(h^5), \\
a_{1,2} = \dfrac{-e^{-\frac{\sqrt{3}}{3}\mu h}\cos\theta + e^{-\frac{3+\sqrt{3}}{3}\mu h}\cos\left(\frac{3-\sqrt{3}}{6}\omega h-\theta\right)}{h\sqrt{\mu^2+\omega^2}\sin\left(\frac{\omega h}{\sqrt{3}}\right)} \\
\quad = \dfrac{3-2\sqrt{3}}{12} + \dfrac{\mu h}{36} - \dfrac{(5\mu^2+\omega^2)h^2}{288\sqrt{3}} + \dfrac{\mu\left(3(7+2\sqrt{3})\mu^2+(29-6\sqrt{3})\omega^2\right)h^3}{12960} \\
\quad - \dfrac{(\mu^2+\omega^2)\left((8+3\sqrt{3})\mu^2-(-4+\sqrt{3})\omega^2\right)h^4}{17280\sqrt{3}} + O(h^5), \\
a_{2,1} = \dfrac{e^{\frac{\sqrt{3}}{3}\mu h}\cos\theta - e^{-\frac{3-\sqrt{3}}{6}\mu h}\cos\left(\frac{3+\sqrt{3}}{3}\omega h-\theta\right)}{h\sqrt{\mu^2+\omega^2}\sin\left(\frac{\omega h}{\sqrt{3}}\right)} \\
\quad = \dfrac{3+2\sqrt{3}}{12} + \dfrac{\mu h}{36} + \dfrac{(5\mu^2+\omega^2)h^2}{288\sqrt{3}} + \dfrac{\left((21-6\sqrt{3})\mu^3+(29+6\sqrt{3})\mu\omega^2\right)h^3}{12960} \\
\quad + \dfrac{(\mu^2+\omega^2)\left((-9+8\sqrt{3})\mu^2+(3+4\sqrt{3})\omega^2\right)h^4}{51840} + O(h^5), \\
a_{2,2} = \dfrac{-\cos\left(\frac{\omega h}{\sqrt{3}}+\theta\right) + e^{\frac{3+\sqrt{3}}{6}\mu h}\cos\left(\frac{3-\sqrt{3}}{6}\omega h+\theta\right)}{h\sqrt{\mu^2+\omega^2}\sin\left(\frac{\omega h}{\sqrt{3}}\right)} \\
\quad = \dfrac{1}{4} - \dfrac{\mu h}{36} - \dfrac{(\mu^2-3\omega^2)h^2}{288\sqrt{3}} + \dfrac{\mu\left(3(3+2\sqrt{3})\mu^2+(1-6\sqrt{3})\omega^2\right)h^3}{12960} \\
\quad - \dfrac{(\mu^2+\omega^2)\left((16+9\sqrt{3})\mu^2-3(4+\sqrt{3})\omega^2\right)h^4}{51840\sqrt{3}} + O(h^5), \\
b_1 = \dfrac{e^{-\frac{3-\sqrt{3}}{6}\mu h}\left(e^{\mu h}\cos\left(\frac{3-\sqrt{3}}{6}\omega h+\theta\right) - \cos\left(\frac{3+\sqrt{3}}{6}\omega h-\theta\right)\right)}{h\sqrt{\mu^2+\omega^2}\sin\left(\frac{\omega h}{\sqrt{3}}\right)} \\
\quad = \dfrac{1}{2} - \dfrac{(\mu^3-\mu\omega^2)h^3}{360\sqrt{3}} + \dfrac{(-3\mu^4-2\mu^2\omega^2+\omega^4)h^4}{8640} + O(h^5), \\
b_2 = \dfrac{e^{-\frac{3+\sqrt{3}}{6}\mu h}\left(-e^{\mu h}\cos\left(\frac{3+\sqrt{3}}{6}\omega h+\theta\right) + \cos\left(\frac{3-\sqrt{3}}{6}\omega h-\theta\right)\right)}{h\sqrt{\mu^2+\omega^2}\sin\left(\frac{\omega h}{\sqrt{3}}\right)} \\
\quad = \dfrac{1}{2} + \dfrac{(\mu^3-\mu\omega^2)h^3}{360\sqrt{3}} + \dfrac{(-3\mu^4-2\mu^2\omega^2+\omega^4)h^4}{8640} + O(h^5),
\end{cases}
\tag{29}
$$

where

$$\theta = \tan^{-1}\left(\frac{\mu}{\omega}\right).$$

Note that the constant terms in the above expansions are just the coefficients of the two-stage Gauss Runge-Kutta method.

Using the method we solve the equation

$$\boldsymbol{y}'(t) = P\,\boldsymbol{y}(t), \quad \boldsymbol{y} = (y_1(t),\, y_2(t),\, y_3(t),\, y_4(t))^T \in \mathbf{R}^4, \tag{30}$$

where

$$P = \begin{pmatrix} -1000 & -10 & 10 & -10 \\ 10 & -1000 & 999 & -1000 \\ 0 & 0 & 0 & -2 \\ 0 & 0 & 1 & -2 \end{pmatrix},$$

and

$$\boldsymbol{y}(0) = (1,\, 2,\, 2,\, 1)^T.$$

This system is stiff since the eigenvalues of coefficient matrix P are given by $-1000 \pm 10\mathrm{i}$ and $-1 \pm \mathrm{i}$. The exact solution is

$$\begin{cases} y_1(t) = e^{-1000\,t}(\cos(10\,t) - \sin(10\,t)), \\ y_2(t) = e^{-t}(\cos t - \sin t) + e^{-1000\,t}(\cos(10\,t) + \sin(10\,t)), \\ y_3(t) = 2\,e^{-t}\cos t, \\ y_4(t) = e^{-t}(\cos t + \sin t). \end{cases}$$

We can see from the solution that the two-stage VCRK with the coefficients evaluated at $\mu = -1$ and $\omega = 1$, say VCRK(1), is expected to be fairly accurate, since the components $e^{-t}\cos t$ and $e^{-t}\sin t$ in the solution become dominant very soon. Moreover, the integration of this stiff system by VCRK(1) is expected to be stable, since the coefficients of the method are close to those of the two-stage Gauss Runge-Kutta, as shown in (29).

Here we integrate the system of the equations by VCRK(1) from $t = 0$ to 5 using the double precision IEEE arithmetic, and compare the result with those of the other two methods: the VCRK with the coefficients evaluated at $\mu = -2$ and $\omega = 2$, say VCRK(2), and the two-stage Gauss Runge-Kutta method. The results are shown in Figure 1.

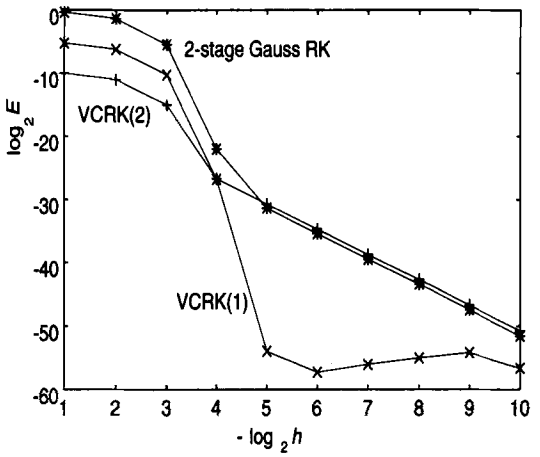

Fig. 1. Global errors at $t = 5$ by the three methods (E is the norm of the error).

We can see from the figure that VCRK(1) is extremely accurate for small h, and that the rate of convergence of VCRK(2) is the same as that of two-stage Gauss Runge-Kutta method, i.e. VCRK(2) is shown to be of order 4, as expected by Theorem 4.

5.2. Three-stage VCRK method

Next we consider the three-stage VCRK with $c_1 = 0$, $c_2 = \frac{1}{2}$ and $c_3 = 1$. For this choice of the abscissae we have

$$\int_0^1 (t - c_1)(t - c_2)(t - c_3)\,\mathrm{d}t = 0, \tag{31}$$

so that $\nu = 1$ in (18). Therefore the order of accuracy of the method is 4, if we use linearly independent functions to determine the coefficients. Here we use the functions

$$\varphi_1(t) = 1, \quad \varphi_2(t) = \sin\omega t, \quad \varphi_3(t) = \cos\omega t, \tag{32}$$

which are linearly independent for $t > 0$, if $\omega \neq 0$. As before we show the closed formulae for the coefficients and their power series expansions:

$$
\begin{cases}
a_{1,1} = a_{1,2} = a_{1,3} = 0 \\
a_{2,1} = \dfrac{\omega h \cos\left(\frac{\omega h}{4}\right) + 2\sin\left(\frac{\omega h}{4}\right) - 2\sin\left(\frac{3\omega h}{4}\right)}{8\omega h \cos\left(\frac{\omega h}{4}\right)\sin^2\left(\frac{\omega h}{4}\right)} \\
\qquad = \dfrac{5}{24} + \dfrac{19\,(\omega h)^2}{5760} + \dfrac{23\,(\omega h)^4}{322560} + O(h^6) \\
a_{2,2} = \dfrac{2\sin\left(\frac{\omega h}{2}\right) - \omega h \cos\left(\frac{\omega h}{2}\right)}{4\omega h \sin^2\left(\frac{\omega h}{4}\right)} = \dfrac{1}{3} - \dfrac{(\omega h)^2}{720} - \dfrac{(\omega h)^4}{80640} + O(h^6) \\
a_{2,3} = \dfrac{\omega h - 4\tan\left(\frac{\omega h}{4}\right)}{8\omega h \sin^2\left(\frac{\omega h}{4}\right)} = -\dfrac{1}{24} - \dfrac{11\,(\omega h)^2}{5760} - \dfrac{19\,(\omega h)^4}{322560} + O(h^6) \\
a_{3,1} = \dfrac{\omega h - 2\sin\left(\frac{\omega h}{2}\right)}{4\omega h \sin^2\left(\frac{\omega h}{4}\right)} = \dfrac{1}{6} + \dfrac{(\omega h)^2}{720} + \dfrac{(\omega h)^4}{80640} + O(h^6) \\
a_{3,2} = \dfrac{2\sin\left(\frac{\omega h}{2}\right) - \omega h \cos\left(\frac{\omega h}{2}\right)}{2\omega h \sin^2\left(\frac{\omega h}{2}\right)} = \dfrac{2}{3} - \dfrac{(\omega h)^2}{360} - \dfrac{(\omega h)^4}{40320} + O(h^6) \\
a_{3,3} = \dfrac{\omega h - 2\sin\left(\frac{\omega h}{2}\right)}{4\omega h \sin^2\left(\frac{\omega h}{4}\right)} = \dfrac{1}{6} + \dfrac{(\omega h)^2}{720} + \dfrac{(\omega h)^4}{80640} + O(h^6) \\
b_1 = a_{3,1}, \quad b_2 = a_{3,2}, \quad b_3 = a_{3,3}
\end{cases}
\tag{33}
$$

In this case, the constant terms in the above expansions are just the coefficients of the three-stage Lobatto IIIA method. Note that this method, like the Lobatto IIIA method, is FSAL(First Same As Last) and therefore effectively two-stage method.

Consider the second order linear ODE

$$y''(t) = -y(t) + \varepsilon \cos t, \tag{34}$$

$$y(0) = 1, \quad y'(0) = 0.$$

The exact solution is given by

$$y(t) = \cos t + \frac{1}{2}\varepsilon\, t\, \sin t.$$

Although the first term of the solution can be integrated exactly by the three-stage VCRK with $\omega = 1$, the second term never be represented exactly unless $\varepsilon = 0$. It is, however, expected that the result of integration by the VCRK with $\omega = 1$ is relatively accurate when ε is small. Here we integrate the equation from $t = 0$ to 10 by the VCRK and three-stage Lobatto IIIA method, which corresponds to the case $h = 0$ in formula (33), for various ε, using the double precision IEEE arithmetic. In Tables 1 and 2, we show the global errors $E = |y_n - y(10)|$, where $nh = 10$, for varying h.

Table 1. Global error $E(\log_2 E)$ of the three-stage VCRK method.

$\log_2 h$	$\varepsilon = 1$	$\varepsilon = 10^{-1}$	$\varepsilon = 10^{-3}$	$\varepsilon = 0$
−1	4.68e−04 (−11.1)	4.68e−05 (−14.4)	4.68e−07 (−21.0)	8.88e−16 (−50.0)
−2	2.94e−05 (−15.1)	2.94e−06 (−18.4)	2.95e−08 (−25.0)	9.99e−16 (−49.8)
−3	1.84e−06 (−19.0)	1.84e−07 (−22.4)	1.84e−09 (−29.0)	5.44e−15 (−47.4)
−4	1.15e−07 (−23.0)	1.15e−08 (−26.4)	1.15e−10 (−33.0)	3.22e−15 (−48.1)
−5	7.21e−09 (−27.0)	7.21e−10 (−30.4)	7.20e−12 (−37.0)	3.55e−15 (−48.0)
−6	4.50e−10 (−31.0)	4.50e−11 (−34.4)	4.28e−13 (−41.1)	2.36e−14 (−45.3)
−7	2.84e−11 (−35.0)	2.94e−12 (−38.3)	1.36e−13 (−42.7)	1.12e−13 (−43.0)
−8	2.07e−12 (−38.8)	3.22e−13 (−41.5)	1.37e−13 (−42.7)	1.23e−13 (−42.9)
−9	9.55e−14 (−43.3)	1.78e−14 (−45.7)	2.37e−14 (−45.3)	2.78e−14 (−45.0)

Table 2. Global error $E(\log_2 E)$ of the three-stage Lobatto IIIA method.

$\log_2 h$	$\varepsilon = 1$	$\varepsilon = 10^{-1}$	$\varepsilon = 10^{-3}$	$\varepsilon = 0$
−1	2.32e−03 (−8.75)	1.87e−04 (−12.4)	4.62e−04 (−11.1)	4.65e−04 (−11.1)
−2	1.46e−04 (−12.7)	1.18e−05 (−16.4)	2.92e−05 (−15.1)	2.94e−05 (−15.1)
−3	9.18e−06 (−16.7)	7.41e−07 (−20.4)	1.83e−06 (−19.1)	1.84e−06 (−19.0)
−4	5.74e−07 (−20.7)	4.63e−08 (−24.4)	1.15e−07 (−23.1)	1.15e−07 (−23.0)
−5	3.59e−08 (−24.7)	2.90e−09 (−28.4)	7.16e−09 (−27.1)	7.21e−09 (−27.0)
−6	2.24e−09 (−28.7)	1.81e−10 (−32.4)	4.48e−10 (−31.1)	4.50e−10 (−31.0)
−7	1.40e−10 (−32.7)	1.13e−11 (−36.4)	2.79e−11 (−35.1)	2.81e−11 (−35.0)
−8	8.63e−12 (−36.8)	7.82e−13 (−40.2)	1.82e−12 (−39.0)	1.83e−12 (−39.0)
−9	1.25e−13 (−42.9)	2.53e−13 (−41.8)	2.97e−13 (−41.6)	2.95e−13 (−41.6)

From Table 1 we can see that the order of accuracy of our VCRK is being 4, and that the method is exact when $\varepsilon = 0$; the values in the column headed $\varepsilon = 0$ undoubtedly show the accumulations of the roundoff errors, since the machine epsilon of the computer is 2.22×10^{-16}. The results of these tables show that the present method is accurate compared with the three-stage Lobatto IIIA method even for the relatively large ε.

6. Embedded Formula

Next we develop an embedded VCRK formula. Let us consider the pair of VCRK formulae $(a_{i,j}, b_i, c_i)$ and $(a_{i,j}, \bar{b}_i, c_i)$, and assume that the stage values $Y_i\,(i = 1, 2, \ldots, s)$, which are common to these two formulae, are r_ith order ap-

proximations to $y(t_n + c_i h)$, i.e.

$$e_i = Y_i - y(t_n + c_i h) = O(h^{r_i+1}), \quad i = 1, 2, \ldots, s. \tag{35}$$

If the coefficients b_i and $\bar{b}_i$ satisfy

$$\begin{cases} B(q) = \sum_{i=1}^{s} b_i \, c_i^{q-1} - \dfrac{1}{q} = O(h^{w+1-q}), & q = 1, 2, \ldots, w, \\ \bar{B}(q) = \sum_{i=1}^{s} \bar{b}_i \, c_i^{q-1} - \dfrac{1}{q} = O(h^{\bar{w}+1-q}), & q = 1, 2, \ldots, \bar{w}, \end{cases} \tag{36}$$

then it is clear from the considerations of the previous section that the local errors E and $\bar{E}$, which correspond to the formulae $(a_{i,j}, b_i, c_i)$ and $(a_{i,j}, \bar{b}_i, c_i)$, respectively, satisfy

$$\begin{cases} E = h f_y \left(\sum_{i=1}^{s} b_i \, e_i \right) + O(h^{w+1}), \\ \bar{E} = h f_y \left(\sum_{i=1}^{s} \bar{b}_i \, e_i \right) + O(h^{\bar{w}+1}). \end{cases} \tag{37}$$

Thus the orders of accuracy of these two methods, say p and $\bar{p}$, are given by

$$p = \min\{r + 1, w\}, \quad \bar{p} = \min\{r + 1, \bar{w}\},$$

$$r = \min_{1 \le i \le s} \{r_i\},$$

where the orders of $\sum_{i=1}^{s} b_i \, e_i$ and $\sum_{i=1}^{s} \bar{b}_i \, e_i$ are assumed to be $r + 1$; we assume that these sums do not annihilate the higher powers in h, unlike the former case. Thus, if we determine the coefficients b_i and $\bar{b}_i$ satisfying the relations

$$w = r + 1, \quad \bar{w} = r,$$

then we have an embedded pair of the methods of orders $p = r + 1$ and $\bar{p} = r$.

As an example, consider the three-stage embedded VCRK formula given by the Butcher array

$$\begin{array}{c|ccc} 0 & & & \\ \frac{1}{2} & a_{2,1} & a_{2,2} & \\ 1 & a_{3,1} & 0 & a_{3,3} \\ \hline & b_1 & b_2 & b_3 \\ \hline & \bar{b}_1 & 0 & \bar{b}_3 \end{array} . \tag{38}$$

The stage and step values of the embedded method are given

$$\begin{cases} y_{n+1} = y_n + h_n(b_1 \, f(t_n, Y_1) + b_2 \, f(t_n + h_n/2, Y_2) + b_3 \, f(t_n + h_n, Y_3)), \\ \bar{y}_{n+1} = y_n + h_n(\bar{b}_1 \, f(t_n, Y_1) + \bar{b}_3 \, f(t_n + h_n, Y_3)), \end{cases} \tag{39}$$

where

$$\begin{cases} Y_1 = y_n, \\ Y_2 = y_n + h_n(a_{2,1}f(t_n, Y_1) + a_{2,2}f(t_n + h_n/2, Y_2)), \\ Y_3 = y_n + h_n(a_{3,1}f(t_n, Y_1) + a_{3,3}f(t_n + h_n, Y_3)). \end{cases}$$

Note that this method is easily parallelizable; we can compute the second and the third stages concurrently on parallel computers. The sets of functions used to determine $(a_{i,j}, \bar{b}_i)$ and b_i are $\{\sin t, \cos t\}$ and $\{t, \sin t, \cos t\}$, respectively. To be specific, we show here the simultaneous equations for these coefficients:

$$\begin{cases} \sin(0.5h) = h(a_{2,1} + a_{2,2}\cos(0.5h)) \\ \cos(0.5h) = 1 - h\,a_{2,2}\sin(0.5h) \\ \sin h = h(a_{3,1} + a_{3,3}\cos h) \\ \cos h = 1 - h\,a_{3,3}\sin h \\ \sin h = h(\bar{b}_1 + \bar{b}_3\cos h) \\ \cos h = 1 - h\,\bar{b}_3\sin h \\ \sin h = h(b_1 + b_2\cos(0.5h) + b_3\cos h) \\ \cos h = 1 - h(b_2\sin(0.5h) + b_3\sin h) \\ h = h(b_1 + b_2 + b_3) \end{cases}$$

In this example $r = 2$, $w = 3$ and $\bar{w} = 2$, and therefore the method of order 2 is embedded in the method of order 3. The stepsize strategy which guarantees the local error of the lower order method within a prescribed tolerance TOL is given by

$$h_{n+1} = \alpha\left(\frac{\mathrm{TOL}}{|y_{n+1} - \bar{y}_{n+1}|}\right)^{1/3} h_n,$$

where α is a safety factor, say $\alpha = 0.9$.

As a numerical example, consider the following two-body problem, a well-known test problem (see [6] and [10]):

$$\begin{cases} y_1'(t) = y_3(t) \\ y_2'(t) = y_4(t) \\ y_3'(t) = -y_1(t)/r^3 \\ y_4'(t) = -y_2(t)/r^3 \end{cases} \tag{40}$$

$$r = \sqrt{y_1^2 + y_2^2},$$

where

$$y_1(0) = 1 - e, \quad y_2(0) = 0, \quad y_3(0) = 0, \quad y_4(0) = \sqrt{(1+e)/(1-e)},$$

and e is an eccentricity. Here we set $e = 0.01$. The numerical solution and the stepsize plot by the method when TOL $= 10^{-5}$ are shown in Figures 2 and 3, and the behavior of the global error is shown in Figure 4.

We can see from Figure 4 that although the global error E increases gradually in average, it remains within the order of 10^{-4}. These figures show that our stepsize strategy controls the local error well.

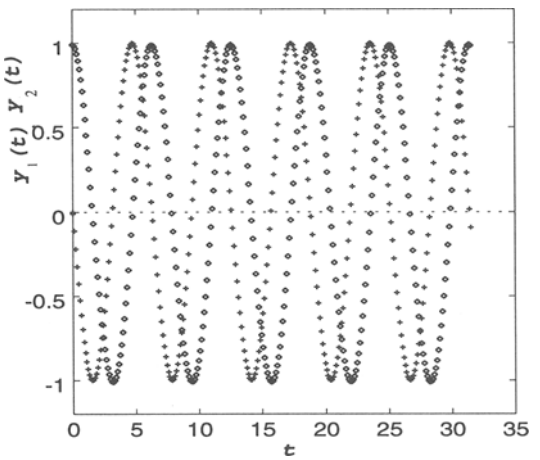

Fig. 2. Numerical solutions $y_{1,n}$ and $y_{2,n}$.

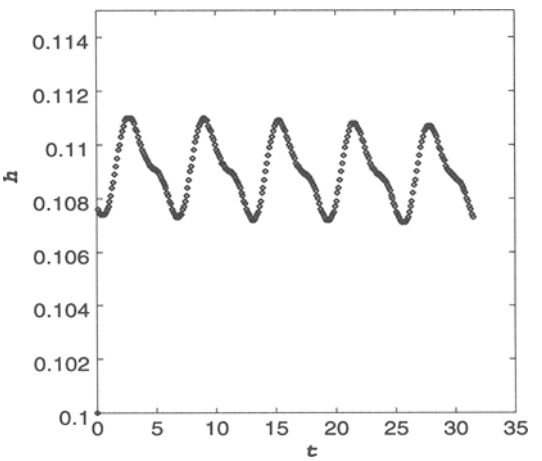

Fig. 3. Stepsize plot.

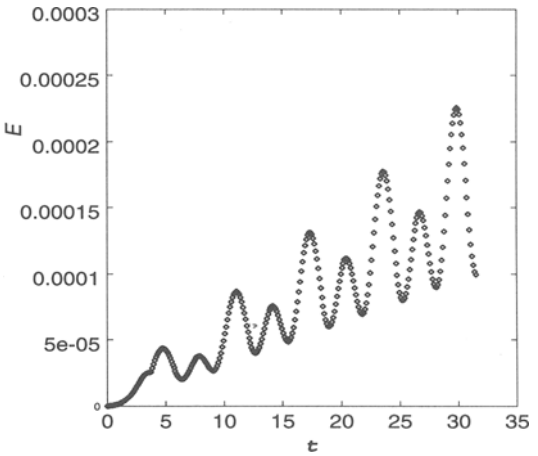

Fig. 4. $E = |y_{1,n} - y_1(t_n)| + |y_{2,n} - y_2(t_n)|$.

7. Conclusion

In this paper we have proposed an s-stage variable coefficient Runge-Kutta (VCRK) method based on the exact integration of s linearly independent functions $\{\varphi_m(t)\}_{m=1}^s$. We have established the existence and uniqueness of the coefficients of the method. It has been shown that the order of accuracy of the method is the same as that of the collocation Runge-Kutta method with the same abscissae c_i. Two- and three-stage methods including an embedded method of this type have been developed. From the numerical experiments, these methods have been shown to be accurate even for the general ODE whose solution cannot be expressed by the linear combination of $\{\varphi_m(t)\}_{m=1}^s$. Stability analysis of the present method will be necessary.

References

[1] D.G. Bettis, Numerical integration of products of Fourier and ordinary polynomials. Numer. Math., **14** (1970), 424–434.
[2] J. Butcher, The Numerical Analysis of Ordinary Differential Equations. Wiley, 1987.
[3] J.P. Coleman, P-stability and exponential-fitting methods for $y'' = f(x, y)$. IMA J. Numer. Anal., **16** (1996), 179–199.
[4] W. Gautschi, Numerical integration of ordinary differential equations based on trigonometric polynomials. Numer. Math., **3** (1961), 381–397.
[5] E. Hairer and G. Wanner, Solving Ordinary Differential Equations II (2nd edition). Springer, 1996.
[6] T.E. Hull, W.H. Enright, B.M. Fellen, and A.E. Sedgwick, Comparing numerical methods for ordinary differential equations. SIAM J. Numer. Anal., **9** (1972), 603–637.
[7] A. Iserles, A First Course in the Numerical Analysis of Differential Equations. Cambridge, 1996.
[8] M. Nakashima, Variable coefficient A-stable explicit Runge-Kutta methods. Japan J. Indust. Appl. Math., **12** (1995), 285–308.
[9] K. Ozawa, A four-stage implicit Runge-Kutta-Nyström method with variable coefficients for solving periodic initial value problems. Japan J. Indust. Appl. Math., **16** (1999), 25–46.

[10] F.L. Shampine, Numerical Solution of Ordinary Differential Equations. Chapman & Hall, 1994.

[11] T.E. Simos, Some new four-step exponential-fitting methods for the numerical solution of the radial Schrödinger equation. IMA J. Numer. Anal., **11** (1991), 347–356.

[12] R.M. Thomas, T.E. Simos and G.V. Mitsou, A family of Numerov type exponential fitted predictor-corrector methods for the numerical integration of the radial Schrödinger equation. J. Comput. Appl. Math., **67** (1996), 255–270.

[13] J. Vanthournout, G. Vanden Berghe and H. De Meyer, Families of backward differentiation methods based on a new type of mixed interpolation. Comput. Math. Appl., **20** (1990), 19–30.