数理解析研究所講究録 1362 巻 2004 年 215-226

Discrete variable methods with variable coefficients for ODEs
常微分方程式における変係数離散変数法

秋田県立大学・システム科学技術学部　小澤　一文 (Kazufumi Ozawa)
Faculty of Systems Science and Technology,
Akita Prefectural University

1 Introduction

A special class of discrete variable methods which is intended to integrate exactly the IVPs with a known solution is derived. This class of the methods, which have variable coefficients, is designed to integrate the ODE exactly only for the case that the solution is a given elementary function, such as trigonometric or exponential function. These methods are expected to be efficient even for the case that the solution is slightly perturbed from the objective functions. A classical example of this class of the methods is the trigonometric linear multistep method of Adams type by Gautschi [5]. This method is designed to be exact, if the solutions are trigonometric functions with a known frequency.

The other examples of this class of methods derived so far are:

1. Störmer and Cowell type trigonometric methods [5], [17].
2. Nyström type trigonometric methods [9].
3. Exponentially fitted linear multistep method [2], [18].
4. Linear multistep method for mixed polynomials [19].
5. Runge–Kutta (–Nyström) type trigonometric methods [10], [15].
6. Runge–Kutta–Nyström method for mixed polynomials [3].

To unify the approaches used to derive the trigonometric and exponential Runge–Kutta(–Nyström) methods, Ozawa [11], [12] has established a technique to adapt the methods to any desired functions (not necessary elementary functions), and has given the condition that the coefficients of such methods exist. He has also established the order conditions for the methods to have order p.

The purpose of this work is to develop a computationally cheap Runge–Kutta method which are exact for a given set of functions, by using the same technique introduced in Ozawa [11], [12].

2 Functionally fitted Runge–Kutta method

Consider the initial value problem

$$y'(t) = f(y(t)), \quad y(0) = y_0, \quad t \in [0, T], \tag{1}$$

and the s-stage Runge–Kutta method

$$\begin{cases} y_{n+1} = y_n + h\sum_{i=1}^{s} b_i f(Y_i), \\ Y_i = y_n + h\sum_{j=1}^{s} a_{i,j} f(Y_j), \qquad i = 1, \ldots, s, \end{cases}$$

for solving the problem (1), where h is a step-size, and y_n is a numerical approximation to the solution $y(t)$ at $t = nh$. Almost all Runge–Kutta methods are designed to be exact when the solution $y(t)$ are polynomials of a given degree or less. In our approach, however, the Runge–Kutta method is designed to be exact not necessary for polynomials but for the linear combinations of predetermined functions $\{\Phi_m(t)\}_{m=1}^s$. We call the functions $\{\Phi_m(t)\}_{m=1}^s$ the *basis functions*, and call the resulting Runge–Kutta method a *functionally fitted Runge–Kutta* (FRK) method.

Here we show a procedure to determine the coefficients of the FRK. First of all, we determine a set of basis functions $\{\Phi_m(t)\}_{m=1}^s$, taking into account the information on the equation or the solution. Next, we give the sparsity pattern of the Butcher array $A = (a_{i,j})$; we consider only the case that the abscissae c_i's are constant and different from each other. In accordance with the sparsity pattern, and with the other requirements (if exist), we set some values (usually 0) to the specified elements of the array. Here we denote by $\mathcal{A}_i$ $(i = 1, \ldots, s+1)$ the set of subscripts of these specified elements in the ith row. Finally, to determine the remaining coefficients $a_{i,j}$ $(j \in \mathcal{A} \setminus \mathcal{A}_i)$, where $\mathcal{A} \equiv \{1, 2, \ldots, s\}$, we choose $(s - |\mathcal{A}_i|)$ different functions from the set of $\Phi_m(t)$'s, and solve the following simultaneous equation:

$$\sum_{j \in \mathcal{A} \setminus \mathcal{A}_i} a_{i,j}\, \Phi_m'(t + c_j\, h) = \frac{\Phi_m(t + c_i\, h) - \Phi_m(t)}{h} - \sum_{j \in \mathcal{A}_i} a_{i,j}\, \Phi_m'(t + c_j\, h), \qquad m \in \mathcal{F}_i \ (i = 1, \ldots, s+1), \tag{2}$$

where we use the convention $a_{s+1,j} = b_j$, and denote by $\mathcal{F}_i \subseteq \mathcal{A}$ the set of the subscripts of the basis functions $\Phi_m(t)$ used in (2). For the uniqueness of the coefficients $a_{i,j}$ and b_j, we assume $|\mathcal{F}_i| = s - |\mathcal{A}_i|$, that is, the number of the unknowns is equal to that of the equations for each i.

For example, suppose we would like to design a three-stage explicit FRK method, then after choosing $\Phi_1(t)$, $\Phi_2(t)$ and $\Phi_3(t)$, we must take $a_{1,1} = a_{1,2} = a_{1,3} = 0$, $a_{2,2} = a_{2,3} = 0$, and $a_{3,3} = 0$, so that

$$\mathcal{A}_1 = \{1, 2, 3\}, \quad \mathcal{A}_2 = \{2, 3\}, \quad \mathcal{A}_3 = \{3\}, \quad \mathcal{A}_4 = \phi,$$
$$\mathcal{F}_1 = \phi, \quad \mathcal{F}_2 = \{1\}, \quad \mathcal{F}_3 = \{1, 2\}, \quad \mathcal{F}_4 = \{1, 2, 3\},$$

and solve the simultaneous equations:

$$a_{2,1}\, \varphi_1(t) = \frac{\Phi_1(t + c_2\, h) - \Phi_1(t)}{h},$$
$$a_{3,1}\, \varphi_m(t) + a_{3,2}\, \varphi_m(t + c_2\, h) = \frac{\Phi_m(t + c_3\, h) - \Phi_m(t)}{h}, \quad m = 1, 2,$$
$$b_1\, \varphi_m(t) + b_2\, \varphi_m(t + c_2\, h) + b_3\, \varphi_m(t + c_3\, h) = \frac{\Phi_m(t + h) - \Phi_m(t)}{h}, \quad m = 1, 2, 3,$$

where $\varphi_m(t) = \Phi_m'(t)$. Note that any choices are possible for the sets $\mathcal{F}_2$ and $\mathcal{F}_3$, only if the conditions $|\mathcal{F}_2| = 1$ and $|\mathcal{F}_3| = 2$ are satisfied. The method obtained in this example is exact for any constant multiple of $\Phi_1(t)$. In general, the method obtained by (2) is exact for the elements of the linear space spanned by the $\Phi_m(t)$'s for $m \in \cap_{i=1}^{s+1} \mathcal{F}_i$, since each stage value Y_i is exact for linear combinations of $\Phi_m(t)$'s for $m \in \mathcal{F}_i$.

The coefficients $a_{i,j}$ and b_i determined in this way depend, in general, not only on h, but also on t. We shall consider, however, the case that these coefficients depend only on h; if the basis functions $\Phi_m(t)$ are polynomials, exponentials or sinusoidal functions, then this is the case, as we will see later. By this assumption, it is possible to take $t = 0$ in (2) without loss of generality.

In [11] and [12], $\mathcal{A}_i = \phi$ and $\mathcal{F}_i = \mathcal{A}$ for all i, that is, there exist s unknowns in each of the simultaneous equations, and all the functions $\Phi_m(t)$ $(m = 1, \ldots, s)$ are used to determine these coefficients. Therefore, the resulting method is necessarily a fully implicit one. For this case, Ozawa [11] has shown that the coefficients given by (2) are unique for all h and $t \in [0, T]$, if the Wronskian matrix associated with $\varphi_m(t) = \Phi'_m(t)$

$$W(t) \equiv \begin{pmatrix} \varphi_1(t) & \cdots & \varphi_s(t) \\ \varphi_1^{(1)}(t) & \cdots & \varphi_s^{(1)}(t) \\ \vdots & \cdots & \vdots \\ \varphi_1^{(s-1)}(t) & \cdots & \varphi_s^{(s-1)}(t) \end{pmatrix}, \tag{3}$$

is nonsingular. Moreover these coefficients are analytic, if all of the functions $\{\Phi_m(t)\}_{m=1}^s$ are analytic on $[0, T]$. Here we extend the result to a general case as follows:

LEMMA 1 *Assume that we are given different constants d_j $(j = 1, \ldots, r)$ and different analytic functions $\psi_m(t)$ $(m = 1, \ldots, r)$. Let $\alpha(h)$ be analytic function at $h = 0$. Then for the given d_k and d_l (not necessarily different), the simultaneous equation*

$$\sum_{j=1}^{r} \alpha_j(h)\, \psi_m(d_j\, h) = \frac{\Psi_m(d_k\, h) - \Psi_m(0)}{h} - \alpha(h)\, \psi_m(d_l\, h), \qquad m = 1, \ldots, r, \tag{4}$$

$$\Psi_m(t) = \int \psi_m(t)\, \mathrm{d}t$$

has unique analytic solutions $\alpha_j(h)$ $(j = 1, \ldots, r)$, if the Wronskian matrix associated with $\psi_m(t)$

$$W_\psi(t) \equiv \begin{pmatrix} \psi_1(t) & \cdots & \psi_r(t) \\ \psi_1^{(1)}(t) & \cdots & \psi_r^{(1)}(t) \\ \vdots & \cdots & \vdots \\ \psi_1^{(r-1)}(t) & \cdots & \psi_r^{(r-1)}(t) \end{pmatrix} \tag{5}$$

is nonsingular.

Although this lemma corresponds to the case that $|\mathcal{A}_i| = 1$ in (2), it is straightforward matter to extend the result to the general case that $|\mathcal{A}_i| \geq 1$.

3 Local truncation error of FRK method

In general, the numerical results given by the FRK will have truncation errors, except for the cases that the method is fitted to the problem (1) completely. Therefore, we must evaluate the errors by using "order of accuracy." The definition of the measure for the FRK is the same as is used for conventional methods. That is, if the numerical solution by the FRK satisfies

$$y_1 - y(h) = \mathrm{O}(h^{p+1}), \quad y(0) = y_0, \quad h \to 0,$$

for any sufficiently smooth solution $y(t)$, then we shall call the integer p the *order of accuracy* of the FRK. However, unlike the conventional case, we must consider the errors in the situation that the coefficients $a_{i,j}$ and b_i also vary as functions of h, when $h \to 0$.

To analyze the local truncation error of the FRK, let us introduce the following quantities:

$$B(q) \equiv \sum_i b_i c_i^{q-1} - \frac{1}{q}, \tag{6}$$

$$C_i(q) \equiv \sum_j a_{i,j} c_j^{q-1} - \frac{c_i^q}{q}, \qquad i = 1, \ldots, s, \tag{7}$$

$$D(q) \equiv \sum_i b_i C_i(q), \tag{8}$$

where $a_{i,j}$ and b_i are the coefficients generated by (2).

In [11] and [12], for the case $\mathcal{A}_i = \phi$, Ozawa has shown

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

For the present case, this result is straightforwardly extended to

$$\begin{aligned} B(q) &= O(h^{r_{s+1}+1-q}), \qquad q = 1, \ldots, r_{s+1}, \\ C_i(q) &= O(h^{r_i+1-q}), \qquad q = 1, \ldots, r_i, \qquad i = 1, \ldots, s. \end{aligned} \tag{9}$$

where we set $r_i = |\mathcal{F}_i|$ $(i = 1, 2, \ldots, s+1)$. We express the errors at the stages and step by using $B(q)$ and $C_i(q)$. First we consider the residuals at the stages and step. Let $y(t)$ be any sufficiently smooth function (not necessary the solution of (1)), then

$$\begin{aligned} R &\equiv y(0) + h \sum_i b_i y'(c_i h) - y(h) = \sum_{q \geq 1} \frac{h^q B(q)}{(q-1)!} (y'(0))^{(q-1)}, \\ R_i &\equiv y(0) + h \sum_j a_{i,j} y'(c_j h) - y(c_i h) = \sum_{q \geq 1} \frac{h^q C_i(q)}{(q-1)!} (y'(0))^{(q-1)}. \end{aligned} \tag{10}$$

Note that if $y(t) = \Phi_m(t)$ these residuals vanish, that is,

$$\begin{aligned} \sum_{q \geq 1} \frac{h^q B(q)}{(q-1)!} (\varphi_m(0))^{(q-1)} &= 0, \qquad m \in \mathcal{F}_{s+1}, \\ \sum_{q \geq 1} \frac{h^q C_i(q)}{(q-1)!} (\varphi_m(0))^{(q-1)} &= 0, \qquad m \in \mathcal{F}_i. \end{aligned} \tag{11}$$

On the other hand, if $\Phi_m(t)$ are polynomials of some degree or less, then $B(q)$ and $C_i(q)$ vanish for the first several q's, and $\varphi_m^{(q-1)}(t) = 0$ for the other higher q's. From (9) and (10) we have

$$R = O(h^{r+1}), \qquad R_i = O(h^{\rho+1}), \tag{12}$$

where

$$\rho = \min_i \{r_i\}, \qquad r = r_{s+1}.$$

Next we consider the relation between the residuals and local errors of the FRK method.

Let $y(t)$ be the solution of $y'(t) = f(y(t))$, then the errors at the stages are given by

$$e_i \equiv Y_i - y(c_i h) = y_0 + h \sum_j a_{i,j} f(Y_j) - \left(y_0 + h \sum_j a_{i,j} y'(c_j h) - R_i \right)$$
$$= h f_y \sum_j a_{i,j} (e_j + O(e_j^2)) + R_i,$$

therefore

$$e_i = (1 - a_{i,i} h f_y)^{-1} \left((h f_y) \sum_{j \neq i} a_{i,j} (e_j + O(e_j^2)) + R_i \right) = O(h^{\rho+1}).$$

For the error at the step, we have

$$E \equiv y_1 - y(h) = y_0 + h \sum_i b_i f(Y_i) - \left(y_0 + h \sum_i b_i y'(c_i h) - R \right)$$
$$= h f_y \sum_i b_i \left(Y_i - y(c_i h) + O(e_i^2) \right) + R. \tag{13}$$

Before evaluating E, we must evaluate the two quantities

$$\sum_i b_i Y_i = \sum_i b_i y_0 + h \sum_{i,j} b_i a_{i,j} f(Y_j),$$
$$\sum_i b_i y(c_i h) = \sum_i b_i y_0 + h \sum_{i,j} b_i a_{i,j} y'(c_j h) - T,$$

where we put

$$T = \sum_i b_i R_i = \sum_{q \geq 1} \frac{h^q D(q)}{(q-1)!} (y'(0))^{(q-1)}. \tag{14}$$

For the order of T, if we assume

$$T = O(h^{\tau+1}), \tag{15}$$

then from (12) we have

$$\tau \geq \rho = \min_i \{r_i\}.$$

Thus

$$E = (h f_y) \sum_{i,j} b_i a_{i,j} \left(f(Y_j) - y'(c_j h) \right) + (h f_y) T + R + O(h^{2\rho+3})$$
$$= (h f_y)^2 \sum_{i,j} b_i a_{i,j} e_j + (h f_y) T + R + O(h^{2\rho+3}).$$

If the order of $\sum_{i,j} b_i a_{i,j} e_j$ is that of the minimum of e_j's, then we have

$$E = O(h^{p+1}),$$

where

$$p = \min \{\rho + 2, \tau + 1, r\}. \tag{16}$$

Thus the order of accuracy of the method is given by (16).

4 Three-stage FESDIRK method

Let us consider the three-stage Runge–Kutta method given by the Butcher array

$$\begin{array}{c|ccc} 0 & 0 & & \\ c_2 & a_{2,1} & \alpha & \\ c_3 & a_{3,1} & a_{3,2} & \alpha \\ \hline & b_1 & b_2 & b_3 \end{array} \tag{17}$$

Usually the methods of this type are called *explicit SDIRK* (ESDIRK) method when the coefficients are constant, and we shall call it *functionally fitted ESDIRK* (FESDIRK) method, if the method is FRK.

For the FESDIRK given by (17), we set

$$\mathcal{A}_1 = \{1, 2, 3\}, \quad \mathcal{A}_2 = \{3\}, \quad \mathcal{A}_3 = \{3\}, \quad \mathcal{A}_4 = \phi,$$
$$\mathcal{F}_1 = \phi, \quad \mathcal{F}_2 = \{1, 2\}, \quad \mathcal{F}_3 = \{1, 2\}, \quad \mathcal{F}_4 = \{1, 2, 3\}.$$

Note that the α in the third row of the array is just the value that has been obtained in the second row so that $|\mathcal{F}_3| = 2$. The simultaneous equations to be solved for these coefficients are

$$\begin{aligned} a_{2,1}\varphi_m(0) + \alpha\,\varphi_m(c_2\,h) &= \frac{\Phi_m(c_2\,h) - \Phi_m(0)}{h}, && m \in \mathcal{F}_2, \\ a_{3,1}\varphi_m(0) + a_{3,2}\,\varphi_m(c_2\,h) &= \frac{\Phi_m(c_3\,h) - \Phi_m(0)}{h} - \alpha\,\varphi_m(c_3\,h), && m \in \mathcal{F}_3, \\ b_1\varphi_m(0) + b_2\,\varphi_m(c_2\,h) + b_3\,\varphi_m(c_3\,h) &= \frac{\Phi_m(h) - \Phi_m(0)}{h}, && m \in \mathcal{F}_4, \end{aligned} \tag{18}$$

where we assume that the Wronskian matrix

$$W(t) = \begin{pmatrix} \varphi_1(t) & \varphi_2(t) & \varphi_3(t) \\ \varphi_1^{(1)}(t) & \varphi_2^{(1)}(t) & \varphi_3^{(1)}(t) \\ \varphi_1^{(2)}(t) & \varphi_2^{(2)}(t) & \varphi_3^{(2)}(t) \end{pmatrix} \tag{19}$$

is nonsingular at $t = 0$. From the construction, it follows that the method is exact when the solution satisfies $y(t) \in \mathrm{span}\{\Phi_1(t), \Phi_2(t)\}$. For this case, we have

$$r_2 = r_3 = 2, \quad r_4 = 3, \quad \rho = 2, \quad \tau \geq 2,$$

and

$$\begin{aligned} B(q) &= \sum_{i=1}^{3} b_i\,c_i^{q-1} - \frac{1}{q} = \mathrm{O}(h^{4-q}), && q = 1, 2, 3, \\ C_i(q) &= \sum_{j=1}^{3} a_{i,j}\,c_j^{q-1} - \frac{c_i^q}{q} = \mathrm{O}(h^{3-q}), && q = 1, 2, \end{aligned} \tag{20}$$

which leads to $p = 3$ from (16).

When $h \to 0$, FESDIRK approaches a constant coefficient method, which has a key role in later considerations. Let $a_{i,j}^{(0)}$ and $b_i^{(0)}$ be the constant terms of the power series expansions of $a_{i,j}$

and b_i, respectively. Then relation (20) means that

$$\sum_{i=1}^{3} b_i^{(0)} c_i^{q-1} = \frac{1}{q}, \qquad q = 1, 2, 3, \tag{21}$$

$$\sum_{j=1}^{i} a_{i,j}^{(0)} c_j^{q-1} = \frac{c_i^q}{q}, \qquad q = 1, 2. \tag{22}$$

The relations (21) and (22), which are the so-called simplifying assumptions [1], determine $a_{i,j}^{(0)}$ and $b_i^{(0)}$ uniquely as functions of c_2. The results are:

$$\begin{cases} a_{2,1}^{(0)} = \dfrac{c_2}{2}, \quad a_{2,2}^{(0)} = \dfrac{c_2}{2} \, (= \alpha), \\ a_{3,1}^{(0)} = -\dfrac{36\,c_2^4 - 120\,c_2^3 + 134\,c_2^2 - 60\,c_2 + 9}{8\,c_2\,(3\,c_2 - 2)^2}, \\ a_{3,2}^{(0)} = -\dfrac{24\,c_2^3 - 50\,c_2^2 + 36\,c_2 - 9}{8\,c_2\,(3\,c_2 - 2)^2}, \quad a_{3,3}^{(0)} = \alpha, \\ b_1^{(0)} = \dfrac{6\,c_2^2 - 6\,c_2 + 1}{6\,c_2\,(4\,c_2 - 3)}, \\ b_2^{(0)} = \dfrac{1}{6\,c_2\,(6\,c_2^2 - 8\,c_2 + 3)}, \\ b_3^{(0)} = \dfrac{2\,(3\,c_2 - 2)^2}{3\,(4\,c_2 - 3)\,(6\,c_2^2 - 8\,c_2 + 3)}. \end{cases}$$

Note that $a_{i,j}^{(0)}$ and $b_i^{(0)}$ are independent of the choice of $\Phi_m(t)$.

5 Fourth order FESDIRK method

We have obtained a three-stage FESDIRK method and have shown that the method is of order 3. To raise the order of the method up to 4 we assume two conditions.

The first condition is

$$\int_0^1 t^{q-1} \cdot t\,(t - c_2)\,(t - c_3)\,\mathrm{d}t \begin{cases} = 0, & q = 1, \\ \neq 0, & q \geq 2. \end{cases}$$

In [14], the case that the integral equals to 0 even for $q \geq 2$ is considered. From this assumption we have

$$c_3 = \frac{4\,c_2 - 3}{2\,(3\,c_2 - 2)}. \tag{23}$$

By assuming (23), we have from [11]

$$B(q) = \sum_{i=1}^{3} b_i\, c_i^{q-1} - \frac{1}{q} = \mathrm{O}(h^{\max\{5-q,\,2\}}), \qquad q = 1, \ldots, 4, \tag{24}$$

so that $r = 4$ in (12), and we have, instead of (21),

$$\sum_{i=1}^{3} b_i^{(0)} c_i^{q-1} = \frac{1}{q}, \qquad q = 1, \ldots, 4, \tag{25}$$

which is the constant term of $B(q)$.

The second assumption is

$$\sum_i b_i^{(0)} a_{i,j}^{(0)} = b_j^{(0)} (1 - c_j), \qquad j = 1, 2, 3. \tag{26}$$

It has been shown that this condition together with (22) and (25) is a sufficient condition for the method $(a_{i,j}^{(0)}, b_i^{(0)}, c_i)$ to be of order 4 (see [1], [6]).

Next lemma shows that conditions (22), (25) and (26) guarantee $\tau = 3$ in (15).

LEMMA 2 *If conditions* (22), (25) *and* (26) *hold, then*

$$D(q) = O(h^{4-q}), \qquad q = 1, 2, 3,$$

so that $\tau = 3$ *in* (15).

Proof. The detail of the proof is shown in [14]. ∎

Since $r = 4$ has already been established, and $\tau = 3$ has been proved by the above lemma, it is clear from (16) that $p = 4$. Thus we have the following theorem:

THEOREM 1 *If the abscissae* c_2 *and* c_3 *satisfy the two conditions* (23) *and* (26), *then FESDIRK with the coefficients given by* (2) *is of order 4.*

Hereafter we call the (F)ESDIRK obtained now (F)ESDIRK4. Next we must obtain the values of c_2 for which condition (26) is valid. Let d_j be

$$d_j = \sum_i b_i^{(0)} a_{i,j}^{(0)} - b_j^{(0)} (1 - c_j), \qquad j = 1, 2, 3,$$

then from (21) and (22) we have

$$\begin{aligned} \sum_j d_j c_j^{q-1} &= \sum_{i,j} b_i^{(0)} a_{i,j}^{(0)} c_j^{q-1} - \sum_j b_j^{(0)} (1 - c_j) c_j^{q-1} \\ &= \frac{1}{q} \sum_i b_i^{(0)} c_i^q - \frac{1}{q} + \frac{1}{q+1} = 0, \qquad \text{for } q = 1, 2, \end{aligned}$$

that is,

$$\begin{aligned} d_1 + \quad d_2 + \quad d_3 &= 0, \\ c_2 d_2 + \; c_3 d_3 &= 0. \end{aligned}$$

This means that if we force one of d_i's to be 0, then the remainders become 0, provided that $0 < c_2 \neq c_3$. Thus we put, for example,

$$d_1 = -\frac{(3\,c_2 - 1)\,(3\,c_2 - 2)\,(c_2 - 1)}{6\,c_2\,(4\,c_2 - 3)} = 0,$$

which leads to

$$c_2 = \frac{1}{3}, \quad \frac{2}{3}, \quad 1.$$

Among these solutions, $c_2 = 2/3$ is not allowed because of (23), so that we consider the remaining two solutions. Comparing the stability regions of the ESDIRK4's with $c_2 = 1/3$ and $c_2 = 1$, and the classical Runge–Kutta method (RK4), we find that the ESDIRK4 with $c_2 = 1/3$ is preferable to the ESDIRK4 with $c_2 = 1$ (see [14]). Therefore we take $c_2 = 1/3$ also for FESDIRK4, since it is expected that FESDIRK has approximately the same properties as those of ESDIRK, when h is small. Hereafter, we simply denote the methods ESDIRK4 and FESDIRK4 with $c_2 = 1/3$, by ESDIRK4 and FESDIRK4, respectively.

6 Numerical example

To see how well FESDIRK4 is fitted to the special problems for which we can find the basis functions successfully, and whether or not the global error of the method behaves like $O(h^4)$ for general problems, we shall present some numerical examples. Here we solve the following three problems:

Airy equation

Constant coefficient linear equation

The solution of the Airy equation oscillates with varying "frequency." The solution of the linear equation consists of the two components: rapidly damped oscillatory component and decaying exponential component. To generate the coefficients of FESDIRK4, we use sinusoidal bases for the Airy equation, and exponential bases for the linear equation. In these experiments, we measure the errors by the Euclidean norms. All the computations are performed by the IEEE double precision arithmetic.

Airy equation

Consider the Airy equation

$$y''(t) - t\,y(t) = 0, \tag{27}$$

with the initial condition

$$\begin{aligned} y(-50) &= \mathrm{Ai}(-50) + 0.5\,\mathrm{Bi}(-50) &&= -2.304564997\cdots \times 10^{-1}, \\ y'(-50) &= \mathrm{Ai}'(-50) + 0.5\,\mathrm{Bi}'(-50) &&= 3.963089871\cdots \times 10^{-1}, \end{aligned}$$

where $\mathrm{Ai}(t)$ and $\mathrm{Bi}(t)$ are Airy's Ai and Bi functions, which are linearly independent solutions of Eq. (27) (see [8]). The exact solution of the problem is

$$y(t) = \mathrm{Ai}(t) + 0.5\,\mathrm{Bi}(t).$$

For this problem, the basis functions

$$\Phi_1(t) = t, \quad \Phi_2(t) = \cos(\omega t), \quad \Phi_3(t) = \sin(\omega t), \tag{28}$$

will be appropriate. For this choice of functions, Wronskian matrix (19) is nonsingular if $\omega \neq 0$. In [14], the coefficients derived by using the functions are shown together with their power series expansions in h; when h is small, it is advantageous to use the expansions rather than the closed forms to avoid the cancellations. We integrate the equation from $t = -50$ to 0, by changing the angular frequency ω by the formula

$$\omega = \sqrt{-t},$$

at every integer point $t = -50, -49, \ldots$. Although the two methods are of the same order, the error of FESDIRK4 is compared favourably with that of ESDIRK4 in Fig. 1.

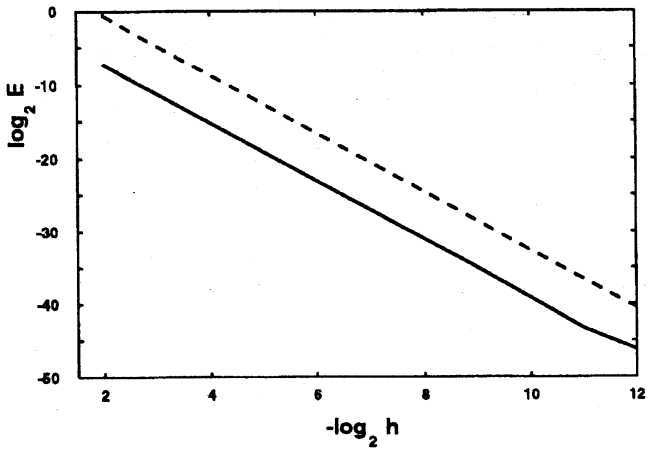

Fig. 1. Errors E of FESDIRK4 (solid) and ESDIRK4 (dashed) versus step-size h for Airy equation (27).

Constant coefficient linear equation

The third problem to be considered is the linear homogeneous equation

$$y'(t) - P\,y(t) = 0, \qquad y(0) = (1, 0, 0, 0)^{\mathrm{T}}, \tag{29}$$

where

$$P = \begin{pmatrix} 0 & 0 & 1 & 101 \\ -96 & -1 & -97 & 6 \\ -98 & 0 & -99 & -96 \\ -1 & 0 & -1 & -102 \end{pmatrix}.$$

The exact solution of the problem is given by

$$y(t) = \begin{pmatrix} \mathrm{e}^{-t} + \mathrm{e}^{-100t} \sin t \\ \mathrm{e}^{-t}(-1+t) + \mathrm{e}^{-100t}(\cos t + 2\sin t) \\ -\mathrm{e}^{-t} + \mathrm{e}^{-100t}(\cos t + \sin t) \\ -\mathrm{e}^{-100t} \sin t \end{pmatrix}.$$

This solution consists of fast and slow modes. If a small step-size which damps out the fast mode is used, then sooner or later the slow mode will dominate the entire solution. Hence, it is advantageous to fit the method to the slow mode rather than the fast mode, when the method is stable. For this reason, we use moderately small step-size and choose the following basis functions:

$$\Phi_1(t) = t, \quad \Phi_2(t) = \exp(-t), \quad \Phi_3(t) = t\exp(-t). \tag{30}$$

The coefficients derived from the functions (30) are shown in [14]. We integrate the equation from $t = 0$ to 2 by the FESDIRK4, and compare the error with those of the three fourth-order Runge–Kutta methods: ESDIRK4, the two-stage Gauss (Gauss2) and the classical Runge–Kutta (RK4) methods. The results are shown in Table 1.

Table 1. Errors of various methods for linear equation (29).

	$\log_2 E$			
$-\log_2 h$	FESDIRK4	ESDIRK4	Gauss2	RK4
2	2.708e+01	2.915e+01	-5.124e+00	1.099e+02
3	2.486e+01	2.713e+01	-2.196e+01	1.531e+02
4	-2.858e+01	-2.585e+01	-2.529e+01	1.682e+02
5	-5.334e+01	-2.985e+01	-2.929e+01	4.702e+01
6	-5.271e+01	-3.387e+01	-3.329e+01	-3.068e+01
7	-5.262e+01	-3.787e+01	-3.729e+01	-3.470e+01
8	-5.125e+01	-4.188e+01	-4.129e+01	-3.870e+01
9	-5.091e+01	-4.586e+01	-4.530e+01	-4.270e+01
10	-5.164e+01	-5.073e+01	-4.907e+01	-4.668e+01
11	-5.212e+01	-5.056e+01	-5.232e+01	-5.163e+01
12	-5.016e+01	-5.062e+01	-4.988e+01	-5.078e+01

E is the Euclidean norm of the error at $t = 2$.

It can been seen that, although FESDIRK4 is less stable than the two-stage Gauss Runge–Kutta method for larger step-sizes, this method is fitted to the solution completely for moderately small step-sizes; the values of order -5.0e+01 or less in the second column of Table 1 are due to the accumulations of round-off errors, since the machine epsilon of the arithmetic is 2^{-53}. On the other hand, although the other methods are not fitted to this problem completely, the errors decrease steadily at the rate of $O(h^4)$, as expected.

References

[1] J. Butcher, *The Numerical Analysis of Ordinary Differential Equations*, Wiley, 1987.

[2] J.R. Cash, A note on the exponential fitting of blended, extended linear multistep methods, BIT **21** (1981), 450–454.

[3] J.P. Coleman and S.C. Duxbury, Mixed collocation methods for $y'' = f(x, y)$, J. Comput. Appl. Math. **126** (2000), 47–75.

[4] J.M. Franco, Embedded pairs of explicit ARKN methods for the numerical integration of perturbed oscillators, Proceedings of the 2002 Conference on Computational and Mathematical Methods on Science and Engineering CMMSE-2002, (Sep. 2002, at Alicante Spain), Vol 1. 92–101.

[5] W. Gautschi, Numerical integration of ordinary differential equations based on trigonometric polynomials, Numer. Math. **3** (1961), 381–397.

[6] E. Hairer, S.P. Nørsett and G. Wanner, *Solving Ordinary Differential Equations I*, Springer-Verlag, Second Revised Edition, 1992.

[7] 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.

[8] N.N. Lebedev, *Special Functions & Their Applications* (Translated & edited by R.A. Silverman), Dover Publications, Inc. 1972.

[9] B. Neta and C.H. Ford, Families of methods for ordinary differential equations based on trigonometric polynomials, J. Computational and Applied Mathematics **10** (1984), 33–38.

[10] K. Ozawa, A four-stage implicit Runge–Kutta–Nyström method with variable coefficients for solving periodic initial value problems, Japan Journal of Industrial and Applied Mathematics, **16** (1999), 25–46.

[11] K. Ozawa, Functional fitting Runge–Kutta method with variable coefficients, Japan Journal of Industrial and Applied Mathematics **18** (2001), 105–128.

[12] K. Ozawa, Functional fitting Runge–Kutta–Nyström method with variable coefficients, Japan Journal of Industrial and Applied Mathematics **19** (2002), 55–85.

[13] K. Ozawa, Functionally fitted linear multistep method, Proceedings of the 2002 Conference on Computational and Mathematical Methods on Science and Engineering CMMSE-2002, (Sep. 2002, at Alicante Spain), Vol 1. 271–280.

[14] K. Ozawa, A functionally fitted three-stage explicit singly diagonally implicit Runge–Kutta method, preprint.

[15] B. Paternoster, Runge–Kutta–Nyström methods for ODEs with periodic solutions based on trigonometric polynomials, Applied Numerical Mathematics **28** (1998), 401–412.

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

[17] E. Stiefel and D.G. Bettis, Stabilization of Cowell's method, Numer. Math. **13** (1969), 154–175.

[18] T.E. Simos, A fourth algebraic order exponentially–fitted Runge–Kutta method for the numerical solution of the Schrödinger equation, IMA J. Numer. Anal. **21** (2001), 919–931.

[19] J. Vanthournout, H. De Meyer and G. Vanden Berghe, Multistep Methods for Ordinary Differential Equations Based on Algebraic and First Order Trigonometric Polynomials, Computational Ordinary Differential Equations (Ed. J.R. Cash and I. Gladwell), 1992, 61–71.

[20] G. Vanden Berghe, H. De Meyer, M. Van Daele and T. Van Hecke, Exponentially–fitted Runge–Kutta methods, Journal of Computational and Applied Mathematics, **125** (2000), 107–115.

[21] P.J. Van Der Houwen and B.P. Sommeijer, Diagonally implicit Runge–Kutta–Nyström methods for oscillatory problems, SIAM J. Numer. Anal. **26** (1989), 414–429.

[22] E.T. Whittaker and G.N. Watson, *A Course of Modern Analysis*, Cambridge University Press, 1973.