A Functionally Fitted Three-stage Explicit Singly Diagonally Implicit Runge-Kutta Method

Kazufumi OZAWA

Faculty of Systems Science and Technology, Akita Prefectural University,
84-4 Tsuchiya-Ebinokuchi, Honjo, Akita 015-0055, Japan
E-mail: ozawa@akita-pu.ac.jp

Received December 1, 2003
Revised March 4, 2005

A special class of Runge-Kutta(-Nyström) methods called functionally fitted (or functional fitting) Runge-Kutta (FRK) methods has recently been proposed by the author. This class of methods is designed to be exact, if the solution of the equation to be solved is an element of the linear space of given functions, which is called the basis functions. The purpose of this article is to develop a functionally fitted Runge-Kutta method that is cheap to implement. The method proposed in this paper is a three-stage explicit singly diagonally implicit Runge-Kutta (ESDIRK) method, which requires one *LU* decomposition per step. This method is exact if the basis functions are properly chosen, and is moderately accurate even if the choice of the functions is inappropriate, since the method is shown to be of order 4 for general cases. An embedded pair of this type is also developed. Several numerical experiments show the superiority of the methods to conventional ones for the particular case that a suitable set of the basis functions can be found.

Key words: functionally fitted Runge-Kutta method, ESDIRK, embedded pair, periodic problem, step-size control

1. Introduction

Initial value problems of ODEs are very important tools in science and technology. When solving the problems, it is often the case that a priori information on the solution and/or equation, such as the period of the solution or the dominant eigenvalue of the coefficient matrix of linear equation, is available. For these cases, if we could design a numerical method based on such information, then the method would be very accurate for the problem. For example, if the solution of the ODE is known to be a sinusoidal function with a small perturbation, then the special method which is exact only for the trigonometric functions with that frequency will be more accurate than general ODE methods.

Many special methods which are exact for trigonometric functions, exponential functions, or mixed-polynomials have been derived (see e.g. [2], [7], [10], [11], [13], [14]). To be able to fit Runge-Kutta (-Nyström) methods to any desired functions, Ozawa [8], [9] has recently developed a technique to construct the Runge-Kutta (-Nyström) method that is exact on the linear space of given functions. When the functions are polynomials, the method reduces to the collocation Runge-Kutta methods. The method proposed by Ozawa, like collocation methods, is fully implicit, so that its computational cost is extremely expensive compared with explicit methods, and expensive with diagonally implicit Runge-Kutta (DIRK) methods.

The purpose of this work is to develop a computationally cheap Runge-Kutta method which are exact for given functions, by using the technique used in Ozawa ([8] and [9]).

2. Functionally Fitted Runge-Kutta Method

Consider the initial value problem

$$\begin{aligned} \frac{\mathrm{d}y(t)}{\mathrm{d}t} &= f(y(t)), \qquad t \in (0, T], \\ y(0) &= y_0, \end{aligned} \tag{1}$$

and the s-stage Runge-Kutta method

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

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:

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

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:

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

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 from (2) is exact for the elements of the linear space spanned by the $\Phi_m(t)$'s for $m \in \bigcap_{i=1}^{s+1} \mathcal{F}_i$, 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. As a result, it is possible to take $t = 0$ in (2) without loss of generality.

In [8] and [9], $\mathcal{A}_i = \phi$ 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 [8] has shown that the coefficients given by using (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 at $t = 0$. 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:

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*

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

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

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 at $t = 0$.

Proof. Consider the matrix given by

$$D(h) \equiv \begin{pmatrix} \psi_1(d_1\,h) & \psi_1(d_2\,h) & \ldots & \psi_1(d_r\,h) \\ \psi_2(d_1\,h) & \psi_2(d_2\,h) & \ldots & \psi_2(d_r\,h) \\ \vdots & \vdots & \cdots & \vdots \\ \psi_r(d_1\,h) & \psi_r(d_2\,h) & \ldots & \psi_r(d_r\,h) \end{pmatrix}.$$

Using the Wronskian matrix $W_\psi(0)$, we can find

$$\begin{aligned} D(h) &= W_\psi^{\mathrm{T}}(0) \cdot \begin{pmatrix} 1 & 1 & \cdots & 1 \\ d_1\,h & d_2\,h & \cdots & d_s\,h \\ \vdots & \vdots & \cdots & \vdots \\ \dfrac{(d_1\,h)^{r-1}}{(r-1)!} & \dfrac{(d_2\,h)^{r-1}}{(r-1)!} & \cdots & \dfrac{(d_r h)^{r-1}}{(r-1)!} \end{pmatrix} + \mathrm{O}(h^r) \\ &= W_\psi^{\mathrm{T}}(0) \cdot \mathrm{diag}\left(1,\, h,\, \ldots,\, \frac{h^{r-1}}{(r-1)!}\right) \cdot V + \mathrm{O}(h^r), \end{aligned} \tag{6}$$

where V is the Vandermonde matrix given by

$$V = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ d_1 & d_2 & \cdots & d_r \\ \vdots & \vdots & \vdots & \vdots \\ d_1^{r-1} & d_2^{r-1} & \cdots & d_r^{r-1} \end{pmatrix}.$$

Since the assumption of this lemma shows that V and $W_\psi(0)$ are nonsingular, $D(h)$ is also nonsingular for small $h > 0$, and therefore α_j are uniquely determined for that case.

Next we consider the limiting case that $h \to 0$. If $D(h) \neq 0$, then we have from Cramer's rule

$$\alpha_j(h) = \frac{\det D_j(h)}{\det D(h)}, \qquad j = 1, \ldots, r, \tag{7}$$

where

$$D_j(h) = \begin{array}{c} \begin{array}{ccccc} & & j & & \end{array} \\ \begin{pmatrix} \psi_1(d_1 h) & \ldots & F_1(h) & \ldots & \psi_1(d_r h) \\ \psi_2(d_1 h) & \ldots & F_2(h) & \ldots & \psi_2(d_r h) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \psi_r(d_1 h) & \ldots & F_r(h) & \ldots & \psi_r(d_r h) \end{pmatrix} \end{array}, \qquad j = 1, \ldots, r,$$

$$F_m(h) \equiv \frac{\Psi_m(d_l h) - \Psi_m(0)}{h} - \alpha(h)\,\psi_m(d_k h), \qquad m = 1, \ldots, r.$$

If we set

$$F_m(h) = \sum_{j \geq 0} \frac{h^j}{j\,!} f_j(h)\,\psi_m^{(j)}(0), \qquad f_j(h) = \frac{d_k^{j+1}}{j+1} - \alpha(h)\,d_l^j,$$

then we have

$$D_j(h) = W_\psi^{\mathrm{T}} \begin{array}{c} \begin{array}{ccccc} \phantom{\frac{(d_1h)^{r-1}}{(r-1)!}} & & j & & \phantom{\frac{(d_rh)^{r-1}}{(r-1)!}} \end{array} \\ \begin{pmatrix} 1 & \ldots & f_0(h) & \ldots & 1 \\ d_1 h & \ldots & f_1(h)\,h & \cdots & d_r h \\ \vdots & \ldots & \vdots & \ldots & \vdots \\ \dfrac{(d_1h)^{r-1}}{(r-1)!} & \ldots & \dfrac{f_{r-1}(h)\,h^{r-1}}{(r-1)!} & \ldots & \dfrac{(d_rh)^{r-1}}{(r-1)!} \end{pmatrix} \end{array} + \mathrm{O}(h^r)$$

$$= W_\psi^{\mathrm{T}} \cdot \operatorname{diag}\left(1,\, h, \ldots, \frac{h^{r-1}}{(r-1)!}\right) \cdot V_j(h) \; + \text{higher-order terms},$$

where

$$V_j(h) = \begin{array}{c} \begin{array}{ccccc} \phantom{d_1^{r-1}} & & j & & \phantom{d_r^{r-1}} \end{array} \\ \begin{pmatrix} 1 & \ldots & f_0(h) & \ldots & 1 \\ d_1 & \ldots & f_1(h) & \ldots & d_r \\ \vdots & \ldots & \vdots & \ldots & \vdots \\ d_1^{r-1} & \ldots & f_{r-1}(h) & \ldots & d_r^{r-1} \end{pmatrix} \end{array},$$

which means that

$$\det D_j(h) = \det W_\psi \cdot \det V_j(h) \cdot \left(\prod_{j=1}^{r} \frac{1}{(j-1)\,!} \right) h^{\frac{r\,(r-1)}{2}} + \text{higher-order terms.}$$

This shows that the orders of $\det D_j(h)$ are at least $h^{\frac{r(r-1)}{2}}$, since $f_j(h) = f_j(0) + \mathrm{O}(h)$, $(h \to 0)$. On the other hand, we have from (6)

$$\det D(h) = \det W_\psi \cdot \det V \cdot \left(\prod_{j=1}^{r} \frac{1}{(j-1)!} \right) h^{\frac{r(r-1)}{2}} + \mathrm{O}(h^{\frac{r(r-1)}{2}+1}).$$

Therefore the singularity at $h = 0$ in (7) is removable, and if we set

$$\alpha_j(0) = \frac{\det V_j(0)}{\det V}, \qquad j = 1, \ldots, r,$$

then α_j are uniquely determined even at $h = 0$. ■

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 some measure. As a measure of the errors we use the "order of accuracy" to evaluate the error. The definition of the measure for the FRK is the same as is used for conventional methods. That is, if the numerical solution given 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:

$$B(q) \equiv \sum_i b_i c_i^{q-1} - \frac{1}{q}, \tag{8}$$

$$C_i(q) \equiv \sum_j a_{i,j} c_j^{q-1} - \frac{c_i^q}{q}, \qquad i = 1, \ldots, s, \tag{9}$$

$$D(q) \equiv \sum_i b_i C_i(q), \tag{10}$$

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

In [8] and [9], for the case $\mathcal{A}_i = \phi$, Ozawa has shown

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

For the present case, this result is straightforwardly extended to

$$\begin{aligned} &B(q) = \mathrm{O}(h^{r_{s+1}+1-q}), && q = 1, \dots, r_{s+1}, \\ &C_i(q) = \mathrm{O}(h^{r_i+1-q}), && q = 1, \dots, r_i, \qquad i = 1, \dots, s, \end{aligned} \tag{11}$$

where we set $r_i = |\mathcal{F}_i|$ $(i = 1, 2, \dots, s+1)$. We express the errors at the stages and step in terms of $B(q)$ 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

$$\begin{aligned} R &\equiv y(0) + h \sum_i b_i\, y'(c_i h) - y(h) = \sum_{q \geq 1} \frac{h^q B(q)}{(q-1)!} (y'(0))^{(q-1)}, \\ R_i &\equiv y(0) + h \sum_j a_{i,j}\, y'(c_j h) - y(c_i h) = \sum_{q \geq 1} \frac{h^q C_i(q)}{(q-1)!} (y'(0))^{(q-1)}. \end{aligned} \tag{12}$$

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

$$\begin{aligned} &\sum_{q \geq 1} \frac{h^q B(q)}{(q-1)!} (\varphi_m(0))^{(q-1)} = 0, && m \in \mathcal{F}_{s+1}, \\ &\sum_{q \geq 1} \frac{h^q C_i(q)}{(q-1)!} (\varphi_m(0))^{(q-1)} = 0, && m \in \mathcal{F}_i. \end{aligned} \tag{13}$$

This is also valid for the case that $\Phi_m(t)$ are polynomials of some degree or less, since 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. In any case, from (11) and (12) we have

$$R = \mathrm{O}(h^{r+1}), \qquad R_i = \mathrm{O}(h^{\rho+1}), \tag{14}$$

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

$$\begin{aligned} e_i &\equiv Y_i - y(c_i h) \\ &= y_0 + h \sum_j a_{i,j}\, f(Y_j) - \Big(y_0 + h \sum_j a_{i,j}\, y'(c_j h) - R_i\Big) \\ &= h f_y \sum_j a_{i,j}\, (e_j + \mathrm{O}(e_j^2)) + R_i, \end{aligned}$$

therefore

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

For the error at the step, we have

$$\begin{aligned} E &\equiv y_1 - y(h) \\ &= y_0 + h \sum_i b_i\, f(Y_i) - \Big(y_0 + h \sum_i b_i\, y'(c_i\, h) - R \Big) \\ &= h \sum_i b_i \left(f(Y_i) - f(y(c_i\, h)) \right) + R \\ &= h\, f_y \sum_i b_i \left(Y_i - y(c_i\, h) + \mathrm{O}(e_i^2) \right) + R. \end{aligned} \tag{15}$$

Before evaluating E, we must evaluate the two quantities

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

where we put

$$T = \sum_i b_i\, R_i = \sum_{q \geq 1} \frac{h^q\, D(q)}{(q-1)!} \, (y'(0))^{(q-1)}. \tag{16}$$

For the order of T, if we assume

$$T = \mathrm{O}(h^{\tau+1}), \tag{17}$$

then from (14) we have

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

Thus

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

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

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

where

$$p = \min \{ \rho + 2, \, \tau + 1, \, r \}. \tag{18}$$

Thus the order of accuracy of the method is given by (18).

4. Three-stage FESDIRK Method

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

$$\begin{array}{c|ccc} 0 & 0 & & \\ c_2 & a_{2,1} & \alpha & \\ c_3 & a_{3,1} & a_{3,2} & \alpha \\ \hline & b_1 & b_2 & b_3 \end{array} \tag{19}$$

Usually the methods of this type are called *explicit singly diagonally implicit Runge-Kutta* (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 (19), 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

$$\begin{aligned} a_{2,1}\varphi_m(0) + \alpha\,\varphi_m(c_2 h) &= \frac{\Phi_m(c_2 h) - \Phi_m(0)}{h}, && m \in \mathcal{F}_2, \\ a_{3,1}\varphi_m(0) + a_{3,2}\,\varphi_m(c_2 h) &= \frac{\Phi_m(c_3 h) - \Phi_m(0)}{h} - \alpha\,\varphi_m(c_3 h), && m \in \mathcal{F}_3, \\ b_1\varphi_m(0) + b_2\,\varphi_m(c_2 h) + b_3\,\varphi_m(c_3 h) &= \frac{\Phi_m(h) - \Phi_m(0)}{h}, && m \in \mathcal{F}_4, \end{aligned} \tag{20}$$

where we assume that the Wronskian matrix

$$W(t) = \begin{pmatrix} \varphi_1(t) & \varphi_2(t) & \varphi_3(t) \\ \varphi_1^{(1)}(t) & \varphi_2^{(1)}(t) & \varphi_3^{(1)}(t) \\ \varphi_1^{(2)}(t) & \varphi_2^{(2)}(t) & \varphi_3^{(2)}(t) \end{pmatrix} \tag{21}$$

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

$$\begin{aligned} B(q) &= \sum_{i=1}^{3} b_i\, c_i^{q-1} - \frac{1}{q} = \mathrm{O}(h^{4-q}), && q = 1,2,3, \\ C_i(q) &= \sum_{j=1}^{3} a_{i,j}\, c_j^{q-1} - \frac{c_i^q}{q} = \mathrm{O}(h^{3-q}), && q = 1,2, \end{aligned} \tag{22}$$

which leads to $p = 3$ from (18).

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 (22) means that

$$\sum_{i=1}^{3} b_i^{(0)} c_i^{q-1} = \frac{1}{q}, \qquad q = 1, 2, 3, \tag{23}$$

$$\sum_{j=1}^{i} a_{i,j}^{(0)} c_j^{q-1} = \frac{c_i^q}{q}, \qquad q = 1, 2. \tag{24}$$

The relations (23) and (24), 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 $\varPhi_m(t)$.

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

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

We will consider the case later that this integral equals to 0 even for $q \geq 2$. From this assumption we have

$$c_3 = \frac{4\,c_2 - 3}{2\,(3\,c_2 - 2)}. \tag{25}$$

Assuming (25), we have from [8]

$$B(q)=\sum_{i=1}^{3} b_i\, c_i^{q-1}-\frac{1}{q}=\mathrm{O}(h^{\max\{5-q,\,2\}}),\qquad q=1,\dots,4, \tag{26}$$

so that $r=4$ in (14), and we have, instead of (23),

$$\sum_{i=1}^{3} b_i^{(0)}\, c_i^{q-1}=\frac{1}{q},\qquad q=1,\dots,4, \tag{27}$$

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

The second assumption is

$$\sum_{i} b_i^{(0)}\, a_{i,j}^{(0)}=b_j^{(0)}\,(1-c_j),\qquad j=1,2,3. \tag{28}$$

It has been shown that this condition together with (24) and (27) is a sufficient condition for the method $(a_{i,j}^{(0)}, b_i^{(0)}, c_i)$ to be of order 4 (see [1], [4]).

Next lemma shows that conditions (24), (27) and (28) guarantee $\tau=3$ in (17).

Lemma 2. *If conditions* (24), (27) *and* (28) *hold, then*

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

so that $\tau=3$ *in* (17).

Proof. Let the power series expansion of $D(q)$ be

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

From the definition of $D(q)$ in (10) and the property of $C_i(q)$ given by (22), we have immediately

$$D(1)=\mathrm{O}(h^2),\qquad D(2)=\mathrm{O}(h),$$

or equivalently

$$D^{(0)}(1)=D^{(1)}(1)=0,$$
$$D^{(0)}(2)=0.$$

Next we show that several terms other than the above vanish. From (24), (27) and (28), we have for $q=1,2,3$

$$D^{(0)}(q)=\sum_i b_i^{(0)}\left(\sum_j a_{i,j}^{(0)}\, c_j^{q-1}-\frac{c_i^q}{q}\right)=\sum_j b_j^{(0)}(1-c_j)\,c_j^{q-1}-\frac{1}{q\,(q+1)}=0. \tag{29}$$

On the other hand, from (16) and (20)

$$\sum_{q\geq 1} \frac{h^q\, D(q)}{(q-1)!}\, (\varphi_m(0))^{(q-1)} = \sum_{\nu\geq 1} \left(\sum_{q=1}^{\nu} \frac{(\varphi_m(0))^{(q-1)}}{(q-1)!}\, D^{(\nu-q)}(q) \right) h^{\nu} \tag{30}$$
$$= 0, \qquad m = 1, 2.$$

Therefore, the condition that the coefficient of h^3 in (30) must be 0 can be written with

$$\begin{aligned} \varphi_1^{(0)}\, D^{(2)}(1) + \varphi_1^{(1)}\, D^{(1)}(2) + \frac{1}{2}\,\varphi_1^{(2)}\, D^{(0)}(3) = 0,\\ \varphi_2^{(0)}\, D^{(2)}(1) + \varphi_2^{(1)}\, D^{(1)}(2) + \frac{1}{2}\,\varphi_2^{(2)}\, D^{(0)}(3) = 0. \end{aligned} \tag{31}$$

Since $D^{(0)}(3) = 0$, which is given by (29), and the submatrix

$$\begin{pmatrix} \varphi_1 & \varphi_2 \\ \varphi_1^{(1)} & \varphi_2^{(1)} \end{pmatrix}$$

of Wronskian matrix (21) is nonsingular by assumption, then

$$D^{(2)}(1) = 0, \qquad D^{(1)}(2) = 0.$$

Summarizing the results obtained so far, we have

$$\begin{aligned} &D^{(0)}(1) = D^{(1)}(1) = D^{(2)}(1) = 0,\\ &D^{(0)}(2) = D^{(1)}(2) = 0,\\ &D^{(0)}(3) = 0. \end{aligned}$$

It is clear from the discussion of this lemma that any other terms of $D^{(l)}(q)$ never vanish. Thus we have proved this lemma. ■

Since $r = 4$ has already been established, and $\tau = 3$ has been proved using the above lemma, it is clear from (18) that $p = 4$. Thus we have the following theorem:

THEOREM 1. *If the abscissae c_2 and c_3 satisfy the two conditions* (25) *and* (28)*, then the 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 (28) 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 (23) and (24) we have

$$\sum_j d_j\, c_j^{q-1} = \sum_{i,j} b_i^{(0)}\, a_{i,j}^{(0)}\, c_j^{q-1} - \sum_j b_j^{(0)}\,(1 - c_j)\, c_j^{q-1}$$

$$= \frac{1}{q}\sum_i b_i^{(0)} c_i^q - \frac{1}{q} + \frac{1}{q+1} = 0, \quad \text{for } q = 1, 2,$$

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 (25), so that we consider the remaining two solutions.

Next we show the stability regions of the ESDIRK4's with $c_2 = 1/3$ and $c_2 = 1$, and compare these regions with that of the classical Runge-Kutta method (RK4).

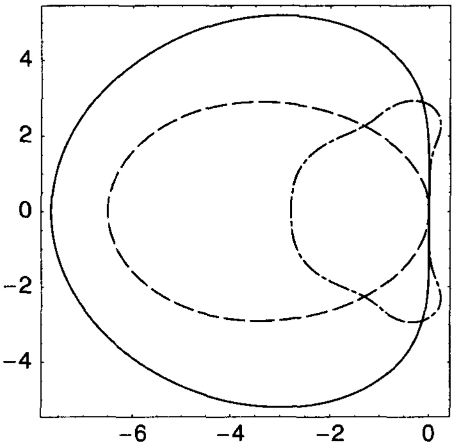

Fig. 1. Stability regions of ESDIRK4's with $c_2 = \frac{1}{3}$ (solid), $c_2 = 1$ (dashed), and RK4 (dash-and-dotted).

Fig. 1 shows that the ESDIRK4 with $c_2 = 1/3$ is preferable to the ESDIRK4 with $c_2 = 1$, since the former has broader stability region. 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. We show the Butcher array of the ESDIRK4 with $c_2 = 1/3$.

$$\begin{array}{c|ccc}
0 & 0 & & \\
\frac{1}{3} & \frac{1}{6} & \frac{1}{6} & \\
\frac{5}{6} & \frac{1}{24} & \frac{5}{8} & \frac{1}{6} \\
\hline
 & \frac{1}{10} & \frac{1}{2} & \frac{2}{5}
\end{array}$$

Hereafter, we simply denote the methods ESDIRK4 and FESDIRK4 with $c_2 = 1/3$, by ESDIRK4 and FESDIRK4, respectively.

Finally we investigate the attainable order with the FESDIRK of the type (19). It is clear from the previous discussion that the FESDIRK and the ESDIRK have always the same order, since the latter corresponds to the particular case that $\Phi_1(t) = t$, $\Phi_2(t) = t^2$, $\Phi_3(t) = t^3$, and the discussion is independent of the choice of the basis functions. Therefore, we will consider the attainable order of the ESDIRK instead of that of the FESDIRK.

There exists unique three-stage six order Runge-Kutta method, that is the Gauss Runge-Kutta method, so that the attainable order of the ESDIRK given by (19) must be at most 5. If the order of the method is 5, then the condition

$$\sum_{i=1}^{3} b_i^{(0)} c_i^{q-1} = \frac{1}{q}, \qquad q = 1, \dots, 5 \tag{32}$$

must be satisfied. For the present case with $c_1 = 0$, the set of the abscissae satisfying this condition is given by

$$c_2 = \frac{6 - \sqrt{6}}{10}, \qquad c_3 = \frac{6 + \sqrt{6}}{10},$$

which is obtained by solving the orthogonality condition

$$\int_0^1 t^q \cdot t\,(t - c_2)\,(t - c_3)\,dt = 0, \qquad q = 0, 1.$$

Substituting the c_2 and c_3 obtained now into (32), and solving this for $b_i^{(0)}$, we have

$$b_1^{(0)} = \frac{1}{9}, \qquad b_2^{(0)} = \frac{16 + \sqrt{6}}{36}, \qquad b_3^{(0)} = \frac{16 - \sqrt{6}}{36}.$$

Moreover, the values of $a_{i,j}^{(0)}$'s which satisfy (24) for these c_i's are given by

$$a_{2,1}^{(0)} = \frac{6 - \sqrt{6}}{20}, \qquad a_{2,2}^{(0)} = \frac{6 - \sqrt{6}}{20},$$

$$a_{3,1}^{(0)} = \frac{6 + \sqrt{6}}{100}, \qquad a_{3,2}^{(0)} = \frac{12 + 7\sqrt{6}}{50}, \qquad a_{3,3}^{(0)} = \frac{6 - \sqrt{6}}{20}.$$

Unfortunately, the set of the values listed above does not satisfy some of the order conditions even for $p = 4$. For example, the order condition corresponding to the tallest tree of order 4, which is given by $\sum_{i,j,k,l} b_i^{(0)} a_{i,j}^{(0)} a_{j,k}^{(0)} a_{k,l}^{(0)} = \frac{1}{24}$, is not satisfied; this becomes $\frac{57-2\sqrt{6}}{1200}$ for the present values. Thus we have the conclusion that the attainable order with the FESDIRK is 4. This means that FESDIRK4 is one of the highest order methods in the class of (19).

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 Problem A, B, C, and D:

Airy equation
Bessel problem
Constant coefficient linear equation
Duffing equation

The solution of Problem A oscillates with varying "frequency." Problems B and D are perturbed oscillators, and the solution of Problem C consists of the two components: rapidly damped oscillatory component and decaying exponential component. To generate the coefficients of FESDIRK4, we use sinusoidal bases for Problems A, B and D, and exponential bases for Problem C. In these experiments, we measure the errors using the Euclidean norms. All the computations are performed by using the IEEE double precision arithmetic.

Airy equation

Consider the Airy equation

$$y''(t) - t\,y(t) = 0, \tag{33}$$

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. (33) (see [6]). The exact solution of the problem is

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

For this problem, the basis functions

$$\Phi_1(t) = t, \quad \Phi_2(t) = \cos(\omega\,t), \quad \Phi_3(t) = \sin(\omega\,t), \tag{34}$$

will be appropriate. For this choice of functions, Wronskian matrix (21) is nonsingular if $\omega \neq 0$. In Appendix A the coefficients derived from 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 will integrate the equation from $t = -50$ to 0, by changing the angular frequency ω with the formula

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

at every integer point $t = -50, -49, \ldots$. The result of Fig. 2 shows that FESDIRK4 is compared favorably with ESDIRK4, although the errors of both methods decrease with the rate of $O(h^4)$.

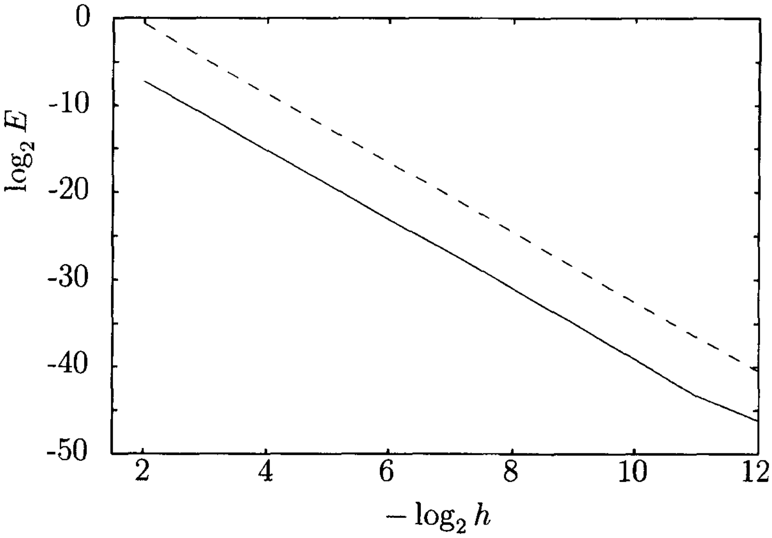

Fig. 2. Errors E of FESDIRK4 (solid) and ESDIRK4 (dashed) versus step-size h for Airy equation (33).

Bessel problem [15]

Next, we consider the equation

$$y''(t) + \left(100 + \frac{1}{4t^2}\right) y(t) = 0, \tag{35}$$

with the initial condition

$$\begin{aligned} y(0.5) &= \frac{1}{\sqrt{2}} J_0(5) &&= -1.255798813\cdots \times 10^{-1}, \\ y'(0.5) &= \frac{1}{\sqrt{2}} J_0(5) - \frac{10}{\sqrt{2}} J_1(5) &&= 2.190754414\cdots, \end{aligned}$$

where $J_\nu(t)$ is the Bessel functions of the first kind. The exact solution of the problem is given by

$$y(t) = \sqrt{t}\, J_0(10\,t).$$

We integrate the equation from $t = 0.5$ to 10 using the two methods: FESDIRK4 through (34) with $\omega = 10$ (fixed) and ESDIRK4. The results are shown in Fig. 3.

From the result, we can observe also in this example that the accuracy of FESDIRK4 is remarkable compared with that of ESDIRK4.

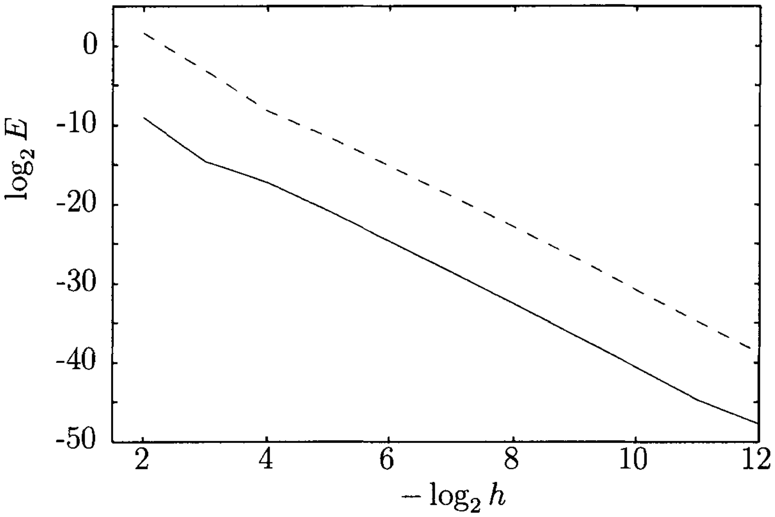

Fig. 3. Errors E of FESDIRK4 (solid) and ESDIRK4 (dashed) versus step-size h for Bessel problem (35).

Constant coefficient linear equation

The third problem to be solved is the linear homogeneous equation

$$y'(t) - Py(t) = 0, \qquad y(0) = (1, 0, 0, 0)^{\mathrm{T}}, \tag{36}$$

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 step-size is within the stability region. For this reason, we use moderately small step-size and choose the following basis functions:

$$\Phi_1(t) = t, \quad \Phi_2(t) = \exp(-t), \quad \Phi_3(t) = t\,\exp(-t). \tag{37}$$

The coefficients derived from the functions (37) are shown in Appendix B. We integrate the equation from $t = 0$ to 2 using 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 (36).

	$\log_2 E$			
$-\log_2 h$	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 the table 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.

Duffing equation [3]

The last equation to be integrated by using FESDIRK4 is a nonlinear equation. Let us consider the Duffing equation

$$y''(t) + (\omega^2 + k^2)\,y(t) - 2\,k^2\,y(t)^3 = 0, \tag{38}$$

$$y(0) = 0, \qquad y'(0) = \omega.$$

The exact solution is given by

$$y(t) = \mathrm{sn}(\omega\,t; (k/\omega)^2),$$

where sn($\cdot$; $\cdot$) is the Jacobian elliptic function. We integrate the equation with $\omega = 1$ and $k = 0.03$ from $t = 0$ to 100 by using FESDIRK4 and ESDIRK4. We use the basis functions (34) with $\omega = 1$, since $\mathrm{sn}(\omega\, t;\ \varepsilon) \to \sin \omega\, t$, as $\varepsilon \to 0$ [16]. The result is shown in Fig. 4.

For this example, although the interval of integration is long compared with those of Problems A and B, these methods give more accurate results, since this equation is subject to small perturbation, which results from the small value of k. Also in this example FESDIRK4 is superior to ESDIRK4.

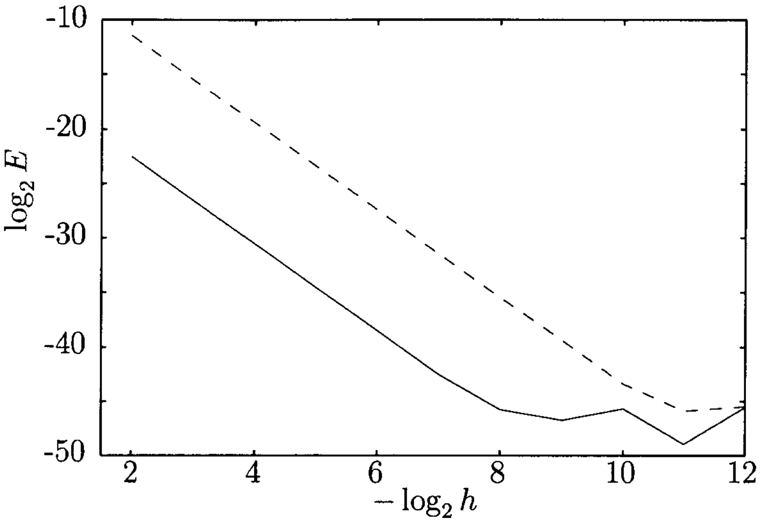

Fig. 4. Errors E of FESDIRK4 (solid) and ESDIRK4 (dashed) versus step-size h for Duffing equation (38).

To summarize, FESDIRK4 is a very efficient scheme for the special problems for which we can find the basis functions successfully. On the other hand, for general problems the method is found to be reasonably accurate since the method is of order 4. The method is not very stable since the method has one explicit stage.

7. Embedded FRK Method

Since we have solved Problems A, B, C, and D, next we must consider 'E' (embedded) method. Let us consider the embedded pair

$$\begin{aligned} y_{n+1} &= y_n + h\ (b_1 f(Y_1) + b_2 f(Y_2) + b_3 f(Y_3)), \\ \bar{y}_{n+1} &= y_n + h\ \left(\bar{b}_1 f(Y_1) + \bar{b}_2 f(Y_2) + \bar{b}_3 f(Y_3) + \bar{b}_4 f(Y_4)\right), \end{aligned} \tag{39}$$

where

$$\begin{cases} Y_1 &= y_n, \\ Y_2 &= y_n + h\,\left(a_{2,1}\,f(Y_1) + \alpha\,f(Y_2)\right), \\ Y_3 &= y_n + h\,\left(a_{3,1}\,f(Y_1) + a_{3,2}\,f(Y_2) + \alpha\,f(Y_3)\right), \\ Y_4 &= y_n + h\,\left(a_{4,1}\,f(Y_1) + a_{4,2}\,f(Y_2) + a_{4,3}\,f(Y_3) + \alpha\,f(Y_4)\right), \end{cases}$$

and we assume $c_2 = 1/3$ and $c_3 = 1$, as before. The Butcher array of the pair is

$$\begin{array}{c|cccc} 0 & 0 & & & \\ \frac{1}{3} & a_{2,1} & \alpha & & \\ \frac{5}{6} & a_{3,1} & a_{3,2} & \alpha & \\ 1 & a_{4,1} & a_{4,2} & a_{4,3} & \alpha \\ \hline & b_1 & b_2 & b_3 & 0 \\ \hline & \bar{b}_1 & \bar{b}_2 & \bar{b}_3 & \bar{b}_4 \end{array}$$

In the array, we further assume

$$\bar{b}_1 = a_{4,1}, \quad \bar{b}_2 = a_{4,2}, \quad \bar{b}_3 = a_{4,3}, \quad \bar{b}_4 = \alpha,$$

so that the method to calculate $\bar{y}_{n+1}$ becomes FSAL (first same as last). The computational cost of this method is approximately the same as that of FESDIRK4, since the number of the LU decomposition to be performed per step is still one.

Here we must determine the coefficients of the method. We take the same $a_{i,j}, b_i\,(1 \le i \le 3)$ and α as those of FESDIRK4, so that the order p of the method corresponds to b_i is 4. With these coefficients, we will determine the $\bar{b}_i$ such that the order of the method which computes $\bar{y}_{n+1}$ is 3.

If we force

$$\bar{B}(q) \equiv \sum_{i=1}^{4} \bar{b}_i\,c_i^{q-1} - \frac{1}{q} = \mathrm{O}(h^{4-q}), \qquad q = 1, 2, 3, \tag{40}$$

then we have

$$\bar{R} \equiv y(0) + h \sum_{i=1}^{4} \bar{b}_i\,y'(c_i h) - y(h) = \sum_{q \ge 1} \frac{h^q\,\bar{B}(q)}{(q-1)!}\,(y'(0))^{(q-1)} = \mathrm{O}(h^4).$$

Therefore, if we set $\bar{R} = \mathrm{O}(h^{\bar{r}+1})$, then $\bar{r} = 3$ and

$$\bar{p} \equiv \min\{\rho + 2,\ \tau + 1,\ \bar{r}\} = \min\{4,\ 4,\ 3\} = 3.$$

The coefficients $\bar{b}_1$, $\bar{b}_2$ and $\bar{b}_3$ satisfying (40) are given by solving the system of the equations

$$\Phi_m(h) = \Phi_m(0) + h\,\left(\bar{b}_1\,\varphi_m(0) + \bar{b}_2\,\varphi_m(c_2\,h) + \bar{b}_3\,\varphi_m(c_3\,h) + \alpha\,\varphi_m(h)\right), \qquad m = 1, 2, 3,$$

for the given $\{\Phi_m(t)\}_{m=1}^3$. With the coefficients, we have the 3rd order method embedded in the 4th order one.

The step-size strategy for this pair, which controls the local truncation error of the lower order method within a prescribed tolerance TOL, is given by

$$h_{n+1} = \theta \left(\frac{TOL}{\|\bar{y}_n - y_n\|} \right)^{1/4} h_n,$$

where θ is a safety factor, say $\theta = 0.9$.

Now, let us apply the embedded pair to the two-body problem [5], [12]

$$y_1'' = -y_1/r^3, \quad y_2'' = -y_2/r^3, \quad r = \sqrt{y_1^2 + y_2^2}, \tag{41}$$

with the initial condition

$$y_1(0) = 1 - e, \;\; y_2(0) = 0, \;\; y_1'(0) = 0, \;\; y_2'(0) = \sqrt{\frac{1+e}{1-e}},$$

where $e \, (0 \le e < 1)$ is an eccentricity. The exact solution of this problem is

$$y_1(t) = \cos u - e, \qquad y_2(t) = \sqrt{1-e^2}\, \sin u,$$

where u is the solution of Kepler's equation

$$u = t + e \sin u.$$

The solution of (41) is found to be 2π-periodic for any e, and is purely sinusoidal for $e = 0$. Hence, a natural choice of the basis functions is (34) with $\omega = 1$. By this choice, the problem with small e is expected to be accurately solved. We integrate the problem with $e = 0.005$ from $t = 0$ to $50\,\pi$ by using the two embedded pairs: FESDIRK4(3) and ESDIRK4(3).

Table 2. Errors and the total steps for two-body problem (41) with $e = 0.005$.

	FESDIRK4(3)		ESDIRK4(3)	
TOL	error	steps	error	steps
10^{-2}	2.785e+00	225	2.483e+00	136
10^{-3}	2.866e-01	170	2.153e+00	277
10^{-4}	7.846e-03	225	1.494e-01	496
10^{-5}	1.399e-03	381	9.359e-03	884
10^{-6}	1.690e-04	680	6.200e-04	1573
10^{-7}	1.846e-05	1207	4.416e-05	2796
10^{-8}	1.938e-06	2144	3.412e-06	4970
10^{-9}	1.993e-07	3806	2.848e-07	8833
10^{-10}	2.021e-08	6762	2.530e-08	15706

TOL: Tolerance of the local error.
error: Euclidean norms of the errors at $t = 50\,\pi$.
steps: Total number of time steps (including rejected steps).

From the result of Table 2, we can see that the embedded method controls the local truncation error well, and as a result, the method integrates the equation with fewer steps compared with ESDIRK4(3).

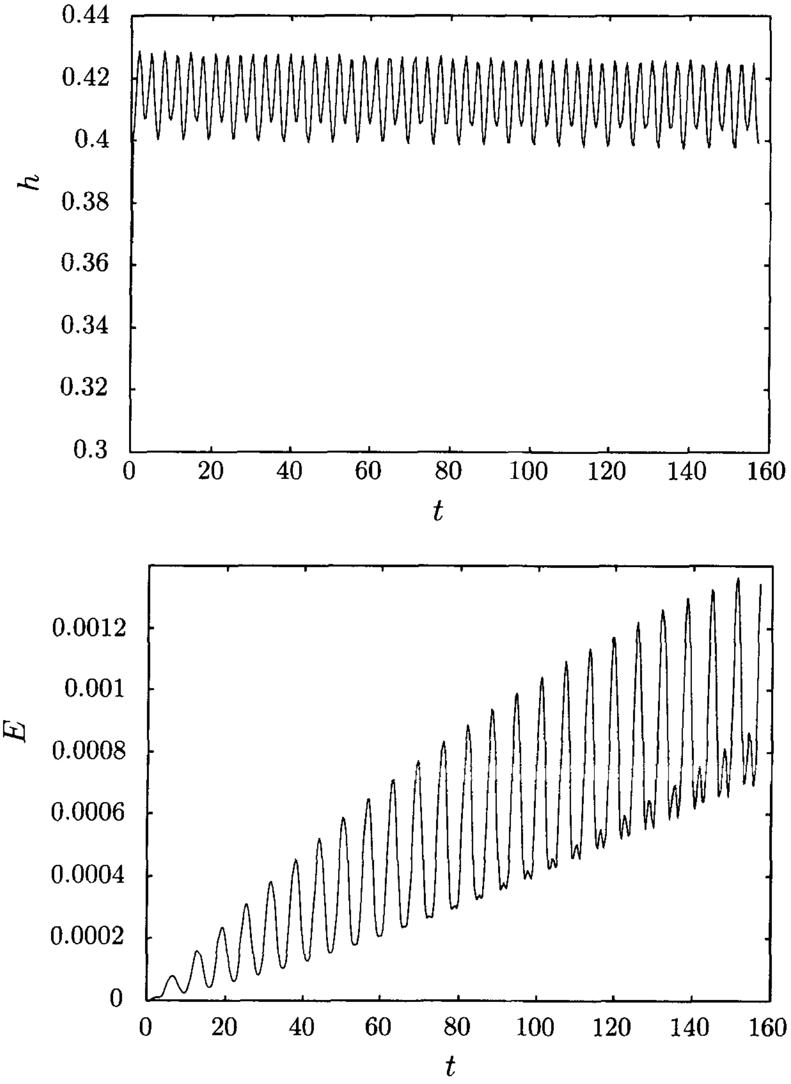

Fig. 5. Step-size plot and error behavior of embedded pair FESDIRK4(3).

8. Summary and Future Work

We have presented a functionally fitted three-stage ESDIRK method. Although the method is of order 4 for general cases, the method is always exact when the solution of the ODE can be expressed in terms of a linear combination of the given basis functions and the method is designed by the functions. Various numerical examples show that the method has proved successful when a suitable set of the basis functions is found, and reasonably accurate, even if this is not the case. The method is extended to an embedded pair.

The stability analysis and the implementation issues of FRKs will be future works.

Acknowledgment. The author wishes to thank Dr. C. Hirota, the research assistant of Akita Prefectural University. He provided much support for this work. The author is also grateful to anonymous referees, whose comments greatly improved this paper.

References

[1] J. Butcher, Numerical Methods for Ordinary Differential Equations. Wiley, 2003.

[2] J.P. Coleman, Mixed interpolation methods with arbitrary nodes. J. Comput. Appl. Math., **92** (1998), 69–83.

[3] 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.

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

[5] 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.

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

[7] 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.

[8] K. Ozawa, Functional fitting Runge-Kutta method with variable coefficients. Japan J. Indust. Appl. Math., **18** (2001), 105–128.

[9] K. Ozawa, Functional fitting Runge-Kutta-Nyström method with variable coefficients. Japan J. Indust. Appl. Math., **19** (2002), 55–85.

[10] 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.

[11] B. Paternoster, Runge-Kutta-Nyström methods for ODEs with periodic solutions based on trigonometric polynomials. Appl. Numer. Math., **28** (1998), 401–412.

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

[13] 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.

[14] G. Vanden Berghe, H. De Meyer, M. Van Daele and T. Van Hecke, Exponentially-fitted Runge-Kutta methods. J. Comput. Appl. Math., **125** (2000), 107–115.

[15] 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.

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

Appendix A.

In the following, we set $\theta = \omega\, h$:

$$
\begin{aligned}
a_{2,1} &= a_{2,2} = a_{3,3} = \frac{\tan(\frac{\theta}{6})}{\theta} = \frac{1}{6} + \frac{\theta^2}{648} + \frac{\theta^4}{58320} + \mathrm{O}(\theta^6),\\
a_{3,1} &= \frac{\sec\left(\frac{\theta}{12}\right)\, \sec^2\left(\frac{\theta}{6}\right)\left(-\sin(\frac{\theta}{4}) + \sin(\frac{5\,\theta}{12})\right)}{4\,\theta}\\
&= \frac{1}{24} - \frac{11\,\theta^2}{10368} - \frac{89\,\theta^4}{3732480} + \mathrm{O}(\theta^6),\\
a_{3,2} &= -\frac{\csc(\frac{\theta}{3})\left(-1 + \cos(\frac{5\,\theta}{6}) + \sin(\frac{5\,\theta}{6})\, \tan(\frac{\theta}{6})\right)}{\theta}\\
&= \frac{5}{8} - \frac{5\,\theta^2}{1152} - \frac{5\,\theta^4}{248832} + \mathrm{O}(\theta^6),\\
b_1 &= \frac{\theta - 2\,\theta\, \cos(\frac{\theta}{6}) + 2\, \sin(\frac{\theta}{2})}{2\,\theta\left(-\cos(\frac{\theta}{3}) + \cos(\frac{\theta}{2})\right)}\\
&= \frac{1}{10} - \frac{\theta^2}{3600} - \frac{163\,\theta^4}{10886400} + \mathrm{O}(\theta^6),\\
b_2 &= \frac{\csc(\frac{\theta}{12})^2\left(\theta\left(-1 + 2\, \cos(\frac{\theta}{6}) - 2\, \cos(\frac{\theta}{3})\right) + 2\, \sin(\frac{\theta}{2})\right)}{4\,\theta\left(1 + 2\, \cos(\frac{\theta}{6})\right)}\\
&= \frac{1}{2} + \frac{\theta^2}{2160} - \frac{61\,\theta^4}{6531840} + \mathrm{O}(\theta^6),\\
b_3 &= -\frac{-\theta\, \cos(\frac{\theta}{6}) + \sin(\frac{\theta}{6}) + \sin(\frac{5\,\theta}{6})}{\theta\left(\cos(\frac{\theta}{6}) - \cos(\frac{2\,\theta}{3})\right)}\\
&= \frac{2}{5} - \frac{\theta^2}{5400} + \frac{397\,\theta^4}{16329600} + \mathrm{O}(\theta^6).
\end{aligned}
$$

Appendix B.

$$
\begin{aligned}
a_{2,1} &= -\frac{3 - 3\,\mathrm{e}^{-\frac{h}{3}} - h}{h^2} = \frac{1}{6} - \frac{h}{54} + \frac{h^2}{648} - \frac{h^3}{9720} + \frac{h^4}{174960} + \mathrm{O}(h^5),\\
a_{2,2} &= a_{3,3} = -\frac{3 - 3\,\mathrm{e}^{\frac{h}{3}} + h}{h^2} = \frac{1}{6} + \frac{h}{54} + \frac{h^2}{648} + \frac{h^3}{9720} + \frac{h^4}{174960} + \mathrm{O}(h^5),\\
a_{3,1} &= -\frac{6 + 3\,\mathrm{e}^{-\frac{5\,h}{6}} - 9\,\mathrm{e}^{-\frac{h}{2}} - 2\,h}{2\,h^2}\\
&= \frac{1}{24} + \frac{11\,h}{216} - \frac{191\,h^2}{10368} + \frac{599\,h^3}{155520} - \frac{6719\,h^4}{11197440} + \mathrm{O}(h^5),
\end{aligned}
$$

$$\begin{aligned}
a_{3,2} &= \frac{\mathrm{e}^{\frac{h}{3}}\left(6+9\,\mathrm{e}^{-\frac{5\,h}{6}}-15\,\mathrm{e}^{-\frac{h}{2}}\right)}{2\,h^2} \\
&= \frac{5}{8}-\frac{5\,h}{72}+\frac{5\,h^2}{384}-\frac{11\,h^3}{10368}+\frac{77\,h^4}{746496}+\mathrm{O}(h^5), \\
b_1 &= \frac{(6-5\,h)\,\mathrm{e}^{-\frac{h}{2}}-(6+h)\,\mathrm{e}^{-\frac{3\,h}{2}}-2\,(3-h)+3\,h^2\,\mathrm{e}^{-\frac{5\,h}{6}}+(6+4\,h)\,\mathrm{e}^{-h}}{\left(2+3\,\mathrm{e}^{-\frac{5\,h}{6}}-5\,\mathrm{e}^{-\frac{h}{2}}\right)h^2} \\
&= \frac{1}{10}+\frac{h^2}{1200}-\frac{13\,h^3}{21600}+\frac{1453\,h^4}{10886400}+\mathrm{O}(h^5), \\
b_2 &= \frac{\mathrm{e}^{\frac{h}{3}}\left(6+\left(-6-6\,(1+h)\;\mathrm{e}^{-\frac{h}{6}}+5\,(1-h)\;h+(6+h)\,\mathrm{e}^{-h}\right)\mathrm{e}^{-\frac{5\,h}{6}}\right)}{\left(2+3\,\mathrm{e}^{-\frac{5\,h}{6}}-5\,\mathrm{e}^{-\frac{h}{2}}\right)h^2} \\
&= \frac{1}{2}-\frac{h^2}{720}+\frac{11\,h^3}{12960}-\frac{781\,h^4}{6531840}+\mathrm{O}(h^5), \\
b_3 &= \frac{\mathrm{e}^{\frac{h}{3}}\left(-6-(6+4\,h)\,\mathrm{e}^{-\frac{4\,h}{3}}+6\,(1+h)\,\mathrm{e}^{-h}+2\,(3-(1-h)\;h)\;\mathrm{e}^{-\frac{h}{3}}\right)}{\left(2+3\,\mathrm{e}^{-\frac{5\,h}{6}}-5\,\mathrm{e}^{-\frac{h}{2}}\right)h^2} \\
&= \frac{2}{5}+\frac{h^2}{1800}-\frac{h^3}{4050}-\frac{227\,h^4}{16329600}+\mathrm{O}(h^5).
\end{aligned}$$