Diagonally Implicit Exponentially Fitted Runge-Kutta Methods with Equation Dependent Coefficients

R. D'Ambrosio and B. Paternoster

Department of Mathematics, University of Salerno, I-84084 Fisciano, Italy

Abstract. It is the purpose of this paper to derive diagonally implicit exponentially fitted methods for the numerical solution of initial value problems based on first order ordinary differential equations. The approach used takes into account the contribution to the error originated from the computation of the internal stages approximations. The derived methods are then compared to those obtained by neglecting the contribution of the error associated to the internal stages, as classically done in the classical derivation of multistage EF-based methods (compare [3] and references therein). Standard and revised EF methods are then compared in terms of linear stability and numerical performances.

Keywords: Runge–Kutta methods, Diagonally Implicit methods, Exponential Fitting.
PACS: 65L05

INTRODUCTION

For a given Hadamard well-posed initial value problems based on first order Ordinary Differential Equations (ODEs)

$$\begin{cases} y' = f(x, y(x)), & x \in [x_0, X] \\ y(x_0) = y_0, \end{cases} \tag{1}$$

we aim to provide an adaptation of Runge-Kutta methods

$$Y_i = y_n + h \sum_{j=1}^{s} a_{ij} f(x_n + c_j h, Y_j),$$
$$y_{n+1} = y_n + h \sum_{i=1}^{s} b_i f(x_n + c_i h, Y_i),$$

following the spirit of the exponential fitting (EF) technique (compare [3] and references therein). Such a technique allows the derivation of numerical methods which are able to accurately integrate problems whose solution exhbits a certain qualitative behaviour which is known a-priori. In particular, in the remainder of this paper, we focus our attention on the family of two-stage singly diagonally-implicit (SDIRK) methods

$$\begin{aligned} Y_1 &= y_n + h\lambda f(x_n + c_1 h, Y_1), \\ Y_2 &= y_n + h\left(a_{21} f(x_n + c_1 h, Y_1) + \lambda f(x_n + c_2 h, Y_2)\right), \\ y_{n+1} &= y_n + h \sum_{i=1}^{2} b_i f(x_n + c_i h, Y_i) \end{aligned} \tag{2}$$

which can be represented in terms of the Butcher tableau

$$\begin{array}{c|cc} c_1 & \lambda & \\ c_2 & a_{21} & \lambda \\ \hline & b_1 & b_2 \end{array}.$$

It is known from the literature that such methods are suitable for an efficient parallel implementation, since the coefficient matrix of the nonlinear system for the computation of the internal stages is lower triangular (see, for instance, [1]). We focus our attention on the derivation of two different EF versions of (2): the first one, in accordance

to the standard EF technique [3], is derived by assuming that the values of the approximations inherited from the computation of the internal stages are exact and, thus, they do not provide any further contribution to the discretization error of the overall scheme; the second version, which follows the spirit of [2, 4], instead takes into account the error provided by the internal stages computation, which cumulates to the truncation error of the overall scheme. Later on, we refer to the former version as *standard* EF-technique, while the latter is denoted as *revised* EF-technique.

STANDARD EF-BASED SDIRK METHODS

We first aim to introduce the family of standard EF-based SDIRK methods (2) applying, as announced, the classical approach presented in [3] and references therein for the derivation of EF multistage methods. To this purpose, we introduce the following linear operators associated to (2)

$$\begin{aligned}
\mathscr{L}_1[h,\mathbf{a}]z(x) &= z(x_n+c_1h) - z(x) - h\lambda z'(x+c_1h),\\
\mathscr{L}_2[h,\mathbf{a}]z(x) &= z(x_n+c_2h) - z(x) - h\left(a_{21}z'(x+c_1h) + \lambda z'(x+c_2h)\right),\\
\mathscr{L}[h,\mathbf{b}]z(x) &= z(x_n+h) - z(x) - h\left(b_1z'(x+c_1h) + b_2z'(x+c_2h)\right),
\end{aligned}$$

where $z(x)$ is assumed to be a smooth enough function. We next define the so-called the fitting space, i.e. the functional basis of the linear space whose elements are the solution of all the problems (1) which are exactly integrated by the EF-method. For the derivation of the standard version of the EF-method (2), we choose the following basis

$$\mathscr{F} = \{1, e^{\mu x}\}, \quad \hat{\mathscr{F}} = \{1, e^{\mu x}, xe^{\mu x}\}, \tag{3}$$

which are respectively associated to the internal and external stages computation: i.e. the internal stages approximations are exact on the linear space spanned by $\mathscr{F}$, while the external stage is exact on the linear space generated by $\hat{\mathscr{F}}$. Thus, we next derive the coefficients λ, a_{21}, b_1 and b_2 by imposing that

$$\mathscr{L}_i[h,\mathbf{a}]z(x) = 0, \quad i = 1,2, \quad \text{for any } z(x) \in \mathscr{F},$$

and

$$\mathscr{L}[h,\mathbf{b}]z(x) = 0, \quad \text{for any } z(x) \in \hat{\mathscr{F}}.$$

We obtain

$$\lambda = \frac{1-e^{-c1z}}{z}, \quad a_{21} = \frac{e^{c_2z}-e^{c_1z}}{ze^{2c_1z}}, \quad b_1 = \frac{1+c_2z+e^z(-1+z-c_2z)}{(c_1-c_2)z^2e^{c_1z}}, \quad b_2 = -\frac{1+c_1z-e^z(1-z+c_1z)}{(c_1-c_2)z^2e^{c_2z}},$$

with $z = \mu h$. We observe that the methods belonging to the derived family have order 2, since

$$\lim_{z\to 0} b_1 + b_2 = 1, \quad \lim_{z\to 0} b_1c_1 + b_2c_2 = \frac{1}{2},$$

which means that the classical conditions of order 2 for Runge-Kutta methods are satisfied when z tends to zero.

REVISED EF-BASED SDIRK METHODS

In standard derivations of EF Runge-Kutta methods, we have computed the coefficients b_1 and b_2 with the underlying assumption that $Y_1 = y(x_n + c_1h)$ and $Y_2 = y(x_n + c_2h)$, i.e. the error in the computation of the internal stages is completely neglected. This is the case only if the solution to the problem (1) is linear combinations of the elements of the functional set $\mathscr{F}$, i.e. 1 and $e^{\mu x}$. Since these two functions are solutions of the differential equation $y'' - \mu y' = 0$, the leading term of the error in the computation of each internal stage is

$$\varepsilon_i = Y_i - y(x_n + c_ih) = h^2F_i(y''(x) - \mu y'(x)), \quad i = 1,2, \tag{4}$$

where F_i is the i-th stage error constant. The stage errors (4) are generally non-zero and, thus, we want to consider their contribution to the error associated to the overall integration process. The knowledge of these errors needs the

calculation of the values of the stage error constants F_i in (4): this is done by following the procedure used in [2], i.e. by solving the linear system

$$\mathscr{L}_i[h,\mathbf{a}]x = \varepsilon_i\Big|_{y(x)=x}, \quad i=1,2,$$

with respect to F_1 and F_2, where ε_i is defined in (4). The obtained values are

$$F_i = \frac{1}{z}\sum_{j=1}^{2} a_{ij} - c_i. \tag{5}$$

We now consider the local error associated to the external stage y_{n+1} in (2)

$$\hat{\mathscr{L}}[h,\mathbf{b}]y(x)\Big|_{x=x_n} = y(x_n+h) - y(x_n) - h\left(b_1^R f(x_n+c_1h, Y_1) + b_2^R f(x_n+c_2h, Y_2)\right), \tag{6}$$

where the superscript R denotes that we are considering *revised* EF methods. Taking into account that

$$y'(x_n+c_ih) = f(x_n+c_ih, Y_i+\varepsilon_i) = f(x_n+c_ih, Y_i) + \varepsilon_i f_y(x_n+c_ih, Y_i) + \mathscr{O}(\varepsilon_i^2) \tag{7}$$

we obtain

$$\hat{\mathscr{L}}[h,\mathbf{b}]y(x)\Big|_{x=x_n} = y(x_n+h) - y(x_n) - h\sum_{i=1}^{2} b_i^R\left(y'(x_n+c_ih) - f_y(x_n+c_ih, Y_i)\varepsilon_i\right).$$

Hereinafter $f_y^{(i)}$ is the short-hand notation for $f_y(x_n+c_ih, Y_i)$. We next evaluate $\hat{\mathscr{L}}^R[h,\mathbf{b}]y(x)$ in correspondence to the elements of $\hat{\mathscr{F}}$ in (3): in particular, we observe that $\hat{\mathscr{L}}^R[h,\mathbf{b}]1$ is automatically equal to zero, while the requested values of $b_1^R(z)$ and $b_2^R(z)$ are those satisfying

$$\hat{\mathscr{L}}^R[h,\mathbf{b}]e^{\mu x}\Big|_{x=0} = \hat{\mathscr{L}}^R[h,\mathbf{b}]xe^{\mu x}\Big|_{x=0} = 0,$$

i.e.

$$b_1^R(z) = \frac{\alpha(z)z^3b_1 + e^{-c_1z}(-1+e^z)f_y^{(2)}h\left(-2e^{c_1z}+e^{c_2z}+e^{2c_1z}(1-c_2z)\right)}{\alpha(z)z^3+\beta(z)hz},$$

$$b_2^R(z) = \frac{\alpha(z)z^3b_2 + (-1+e^z)f_y^{(1)}h\left(1+e^{c_1z}(-1+c_1z)\right)}{\alpha(z)z^3+\beta(z)hz},$$

where

$$\alpha(z) = (c_1-c_2)e^{(2c_1+c_2)z},$$

$$\beta(z) = \left(-2e^{c_1z}f_y^{(2)} + e^{c_2z}\left(f_y^{(1)}+f_y^{(2)}\right) + e^{(c_1+c_2)z}f_y^{(1)}(-1+c_1z) + e^{2c_1z}f_y^{(2)}(1-c_2z)\right).$$

LINEAR STABILITY ANALYSIS

We next consider the linear stability analysis of the derived methods with respect to the classical test problem $y' = \omega y$, with $\mathrm{Re}(\omega) < 0$. The application of the SDIRK method to such problem leads to the recurrence relation $y_{n+1} = R(\nu,z)y_n$, where

$$R(\nu,z) = 1 + \nu\mathbf{b}^T(z)(I-\nu\mathbf{A}(z))^{-1}\mathbf{e},$$

is the stability function of the method, being $\mathbf{e}\in\mathbb{R}^s$ the unit vector. In correspondence to this notion, we recall the following definition of stability region [2].

Definition 1 *The region of the three-dimensional* $(\mathrm{Re}(\nu),\mathrm{Im}(\nu),z)$ *space on which the inequality*

$$|R(\nu,z)| < 1 \tag{8}$$

is satisfied is called a region of stability Ω *for the EF-based method* (2).

Fig. 1 presents a selection of sections through the stability region by planes $z = const$ for fixed values of the nodes $c_1 = 0$ and $c_2 = 1$: we can advise from the picture that the regions corresponding to revised EF methods are larger than those corresponding to standard EF methods. Such a behaviour is more and more visible when the exponential behaviour of the solution becomes more prominent.

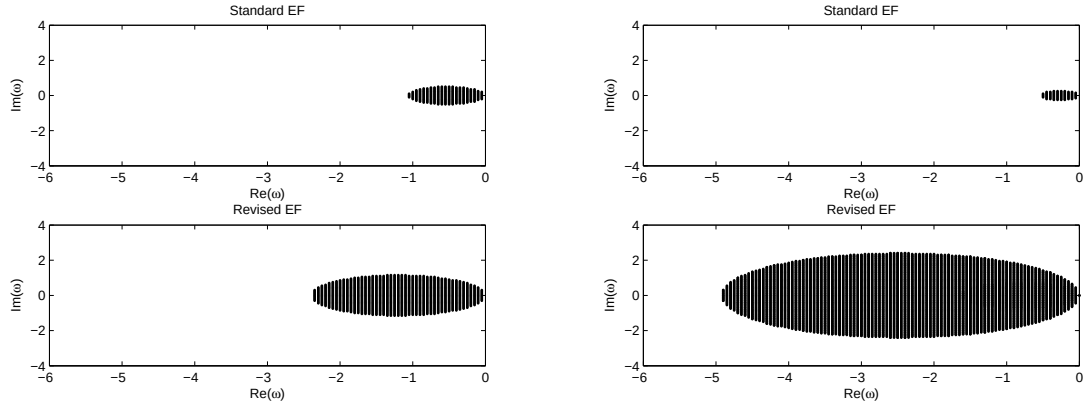

FIGURE 1. Section through the stability regions by planes $z=-2$ (left) and $z=-4$ (right) for (2), with $c_1=0$, $c_2=1$

NUMERICAL EXPERIMENTS

We finally provide a numerical evidence to assert the effectiveness of our approach, by considering the numerical solution of the nonlinear problem

$$y'(x)=\frac{\lambda y^2(x)+2x^3e^{2\lambda x}}{y(x)},\quad y(1)=e^{\lambda}, \tag{9}$$

with $x\in[1,5]$, whose exact solution is $y(x)=x^2e^{\lambda x}\notin\mathscr{F}\cap\hat{\mathscr{F}}$, hence neither the standard EF-based SDIRK method nor the revised one are able to exactly solve it. The computations have been done on a node with CPU Intel Xeon 6 core X5690 3,46GHz, belonging to the E4 multi-GPU cluster of Mathematics Department of Salerno University. The results, reported in Table 1, suggest that, by integrating both methods with the same constant stepsize h, the revised EF one is more accurate.

TABLE 1. Performance of the two versions for the problem (9) for $\lambda=-2$. *err* is the global error achieved in the last integration point, *cd* is the number of gained correct digits, *p* is an estimate to the order of convergence of the method

		Standard EF-SDIRK			Revised EF-SDIRK		
	h	*err*	*cd*	*p*	*err*	*cd*	*p*
$\lambda=-2$	1/ 32	5.82(-04)	3.23		1.24(-04)	3.90	
	1/64	1.42(-04)	3.85	2.03	3.21(-05)	4.49	1.95
	1/128	3.51(-05)	4.45	2.02	8.13(-06)	5.09	1.98
	1/256	8.72(-06)	5.06	2.01	2.05(-06)	5.69	1.99
	1/512	2.17(-06)	5.66	2.01	5.14(-07)	6.29	2.00

ACKNOWLEDGMENTS

The authors express their gratitude to prof. Liviu Gr. Ixaru for the profitable discussions we had on the topic, which have promoted and inspired our advances in the research on EF-based numerical methods.

REFERENCES

1. J. C. Butcher, *Numerical methods for Ordinary Differential Equations, Second Edition*, Wiley, Chichester, 2008.
2. R. D'Ambrosio, L. Gr. Ixaru, B. Paternoster, *Comput. Phys. Commun.* **182**(2), 322–329 (2011).
3. L. Gr. Ixaru, G. Vanden Berghe, *Exponential Fitting*, Kluwer Academic Publishers, Dordrecht, 2004.
4. L. Gr. Ixaru, *Comput. Phys. Commun.*, **183**(1), 63–69 (2012).