J. KSIAM Vol.17, No.4, 238–266, 2013 http://dx.doi.org/10.12941/jksiam.2013.17.238

HIGH ORDER EMBEDDED RUNGE-KUTTA SCHEME FOR ADAPTIVE STEP-SIZE CONTROL IN THE INTERACTION PICTURE METHOD

STÉPHANE BALAC

UEB, UNIVERSITÉ EUROPÉENNE DE BRETAGNE, UNIVERSITÉ DE RENNES I, FRANCE
CNRS, UMR 6082 FOTON, ENSSAT, 6 RUE DE KERAMPONT, CS 80518, 22305 LANNION, FRANCE
E-mail address: stephane.balac@univ-rennes1.fr

ABSTRACT. The Interaction Picture (IP) method is a valuable alternative to Split-step methods for solving certain types of partial differential equations such as the nonlinear Schrödinger equation or the Gross-Pitaevskii equation. Although very similar to the Symmetric Split-step (SS) method in its inner computational structure, the IP method results from a change of unknown and therefore do not involve approximation such as the one resulting from the use of a splitting formula. In its standard form the IP method such as the SS method is used in conjunction with the classical 4th order Runge-Kutta (RK) scheme. However it appears to be relevant to look for RK scheme of higher order so as to improve the accuracy of the IP method. In this paper we investigate 5th order Embedded Runge-Kutta schemes suited to be used in conjunction with the IP method and designed to deliver a local error estimation for adaptive step size control.

1. INTRODUCTION

The *Interaction Picture* (IP) method is a very promising alternative to Split-step methods for solving certain type of partial differential equations (PDE) such as the Gross-Pitaevskii equation (GPE) or the generalised nonlinear Schrödinger equation (GNLSE). The *fourth-order Runge-Kutta method in the Interaction Picture* (RK4-IP) method has been developed by the *Bose-Einstein Condensate Theory Group* of R. Ballagh from the Jack Dodd Centre at the University of Otago in the 90's for solving the Gross-Pitaevskii equation which is ubiquitous in Bose condensation. To our knowledge, it was first described in the Ph.D. thesis of B.M. Caradoc-Davies [1] and M.J. Davis [2]. Since, the RK4-IP method has been widely used for numerical studies concerning Bose-Einstein condensates, see e.g. [3, 4, 5], as well as for numerical simulation of light propagation in optical fibers, see e.g. [6, 7, 8]. The IP method is very similar to the Symmetric Split-step (SS) method in its inner computational structure, since in both cases the method consists in solving in a sequential order over a discretisation grid, one linear PDE problem, one nonlinear ordinary differential equation (ODE) problem, and another linear PDE problem. However the IP method results from a change of unknown and therefore

Received by the editors June 17 2013; Accepted October 15 2013.
2000 *Mathematics Subject Classification.* 78-04, 78M25, 65L06, 35Q60.
Key words and phrases. Interaction Picture method, Embedded Runge-Kutta method, Split-Step method, Gross-Pitaevskii equation, Generalized nonlinear Schrödinger equation.

do not involve approximation – such as the one resulting from the use of a splitting formula in Split-step methods – when obtaining the sequence of the 3 above mentioned problems over each computational step. When the IP method is applied for solving the GP or GNLS equations, the 2 linear PDE problems can be considered to be solved exactly by using the Fourier Transform tools (actually they are solved numerically with high accuracy by using the FFT method). Therefore, the only approximation in the method is introduced by the numerical scheme (typically a Runge-Kutta scheme) required for solving the nonlinear ODE problem. In its standard form the IP method is used in conjunction with the classical 4th order RK scheme because it provides a good compromise between accuracy of the results and complexity of the algorithm of the RK4-IP method and therefore offers a good compromise between accuracy and computation time. However, since the only approximation in the IP method results from the use of a RK scheme for solving the ODE problem it seems to be relevant to look for RK schemes of higher order to improve the accuracy of the results in the IP method. Such RK schemes of high order are numerous in the literature (see e.g. [9, 10, 11]) but unfortunately they have not been designed in order to be used in conjunction with the IP method and therefore are not optimal in this context. Our goal in this paper is to build a RK scheme of order 5 with the constrain of finding the best possible RK coefficients in order to reduce as much as possible the over-cost of the method compared to the standard RK4-IP method.

The name "Interaction Picture" and the change of unknown at the heart of the method originate from quantum mechanics [12, 13] where it is usual to chose an appropriate "picture" in which the physical properties of the studied system can be easily revealed and the calculation made simpler. The classical pictures in quantum mechanics are the Schrödinger and the Heisenberg pictures. The interaction picture is considered as an intermediate between the Schrödinger picture and the Heisenberg picture. It is useful e.g. in quantum optics for solving problems with time-dependent Hamiltonians in which the Hamiltonian can be partitioned as $H(t) = H_0 + V(t)$ where H_0 is a Hamiltonian independent of time and its eigenvalues are easy to compute whereas V is a time-dependent potential which can be complicated. In a numerical context the "Interaction Picture" approach is a way of solving certain PDE involving typically a linear term (e.g. for the GPE the effect of diffusion which links points spatially) and a nonlinear term (e.g. for the GPE the non-diffusive terms which act only locally) that consists in separating the way the 2 groups of terms act in order to solve a much simpler equation. Usually it allows to solve the simpler equation as if it were an ODE by mean of numerical methods for ODE such as Runge-Kutta methods. Actually the choice of operator splitting one should use depends solely on a particular application and no general method is known. The use of the "Interaction Picture" amounts from a mathematical point of view to a change of unknown. The linear part of the equation is handled by the definition of the new unknown itself whereas the nonlinear part remains in the simpler equation to be solved.

In [14] we have proposed an adaptive step-size control version of the RK4-IP method. Namely, we have obtained a 5 stage 3rd order RK formula embedding the standard 4th order RK formula (termed ERK4(3) scheme) with the same features than the standard 4th order RK formula when used in conjunction with the IP method. In particular this ERK4(3) scheme preserves the ease of implementation and the advantageous position of the internal quadrature

nodes of the RK4 formula for the IP method and delivers a local error estimate at no significant extra cost. In this paper we present a higher order embedded RK scheme, namely a 7 stage 4th order RK formula embedding a 5th order RK formula (ERK5(4) scheme) to be used in conjunction with the IP method for adaptive step-size control purposes. One reason for looking for high order RK formulae is that so as to attain a certain accuracy of the results they require less computational steps and therefore are likely to reduce the accumulation of round-off errors. Of course, one single step of a higher order RK formula requires more computations than one step of a lower order RK formula but altogether the higher order RK formula should involved less computations for a better accuracy. There exists in the literature lots of ERK5(4) schemes [9, 10, 11]. However each of these schemes has been constructed in order to satisfy one given criterion and none of them preserve the advantageous position of the internal quadrature nodes of the RK4 formula liable for the efficiency of the ERK4(3)-IP method. In particular, in [15] S.N. Papakostas and G. Papageorgiou propound an algorithm for obtaining a large family of ERK5(4) schemes depending on 5 free parameters under the assumption that the elementary quadrature nodes of the RK pairs are distinct (such RK schemes are termed *quadrature non-defective methods*). This assumption on quadrature nodes is not satisfactory in our quest for a ERK5(4) scheme designed for the IP method. We can mention as well that a Cash-Karp ERK5(4) scheme [16] was used in [2] for solving the Gross-Pitaevskii equation by the IP method but without giving the desired efficiency according to the author. Thus our goal is to construct an embedded RK pair of order 5 and 4 well suited to be used in conjunction with the IP method and preserving the nice features of the ERK4(3)-IP method as far as we can.

The paper is organised as follows. In Section 2 we present an overview of the IP method. Section 3 is devoted to the construction of our embedded RK pair of order 5 and 4 from the general set of order condition equations for RK formulae and to a discussion on the best choice for the free coefficients to optimize the ERK scheme in the context of the IP method. An algorithm for the corresponding ERK5(4)-IP method is also given. In Section 4 we present numerical simulation results in order to illustrate the features of the ERK5(4)-IP method.

2. Overview of the Interaction Picture method

2.1. **PDE problem setting.** We first present a brief summary of the IP method for a general evolution equation in the form of

$$\frac{\partial}{\partial s}u(s,r) = \mathcal{D}\,u(s,r) + \mathcal{N}(u)(s,r), \tag{2.1}$$

where $\mathcal{D}$ and $\mathcal{N}$ denote respectively linear and nonlinear operators (that usually do not commute to each other); the linear differential operator $\mathcal{D}$ includes all the derivation terms with respect to the variable r but does not involve derivation with respect to s and the nonlinear operator $\mathcal{N}$ does not involve derivation at all. This PDE is to be solved for the unknown u in a set $I \times \Omega$ where typically Ω is an open subset in $\mathbb{R}^d$, $d \in \mathbb{N}^*$, and I is an open interval in $\mathbb{R}$. Together with equation (2.1) we consider the initial condition: $u(s=0,r) = \nu_0(r)$, $\forall r \in \Omega$ where ν_0 is a sufficiently regular function from Ω to $\mathbb{C}$.

For instance, for the cubic nonlinear Schrödinger equation

$$\begin{cases} \dfrac{\partial}{\partial t}u(t,\mathbf{r}) + \mathrm{i}\Delta u(t,\mathbf{r}) + \mathrm{i}\epsilon|u(t,\mathbf{r})|^2 u(t,\mathbf{r}) = 0 \quad \forall \mathbf{r} \in \mathbb{R}^2 \; \forall t \in \mathbb{R} \\ u(t=0,\mathbf{r}) = u_0(\mathbf{r}) \quad \forall \mathbf{r} \in \mathbb{R}^2 \end{cases}$$

where $\epsilon = \pm 1$ and Δ stands for the Laplacian operator in $\mathbb{R}^2$, we have $\mathcal{D} : u \mapsto \mathrm{i}\Delta u$ and $\mathcal{N} : u \mapsto \mathrm{i}\epsilon|u|^2 u$. For the generalised nonlinear Schrödinger equation (GNLSE) in optics [17, 8] we are interested in solving the following problem

$$\begin{cases} \dfrac{\partial}{\partial z}A(z,t) = \mathcal{D}A(z,t) + \mathcal{N}(A)(z,t) \quad \forall z \in]0, L[\; \forall t \in \mathbb{R} \\ A(0,t) = a_0(t) \quad \forall t \in \mathbb{R} \end{cases} \tag{2.2}$$

where the unknown A corresponding to the slowly varying optical pulse envelope is a function of time t and position z along the fiber; the linear operator $\mathcal{D}$ is given by

$$\mathcal{D} : A \longmapsto -\frac{1}{2}\alpha A - \sum_{n=2}^{n_{\max}} \beta_n \frac{\mathrm{i}^{n-1}}{n!}\frac{\partial^n}{\partial t^n}A, \tag{2.3}$$

where α is the linear attenuation coefficient of the fiber and $\beta_n, n \geq 2$ are the linear dispersion coefficients of the fiber; the nonlinear operator $\mathcal{N}$ is given by

$$\begin{aligned} \mathcal{N} : A \longmapsto \mathrm{i}\gamma \left[\mathrm{Id} + \frac{\mathrm{i}}{\omega_0}\frac{\partial}{\partial t}\right] \Big(& (1 - f_R)\, A\, |A|^2 \\ & + f_R\, A \int_0^\infty h_R(s)\, |A(\cdot, \cdot - s)|^2 \, \mathrm{d}s \Big), \end{aligned} \tag{2.4}$$

where Id denotes the identity operator, h_R is the Raman time response function, f_R represents the fractional contribution of the delayed Raman response to nonlinear polarisation, γ is the nonlinear fiber parameter and ω_0 is the pulsation of the optical pulse assumed to be quasi-monochromatic. We may notice that another splitting is possible for the GNLSE: the term $-\frac{1}{2}\alpha A$ can be added to the nonlinear operator $\mathcal{N}$ instead of the linear operator $\mathcal{D}$.

For the Gross-Pitaevskii equation (GPE) used to explore the dynamics of vortexes in Bose-Einstein condensates in 2 or 3 space dimensions [1, 2, 18], the condensate wave function ψ is given in the domain Ω occupied by the condensate by

$$\frac{\partial}{\partial t}\psi(\mathbf{r},t) = \mathrm{i}\,\Delta\,\psi(\mathbf{r},t) + \mathcal{N}(\psi)(\mathbf{r},t), \tag{2.5}$$

where Δ is the Laplacian operator in 2 or 3 dimension, and

$$\mathcal{N} : \psi \longmapsto -\mathrm{i}(V\psi + C|\psi|^2\,\psi), \tag{2.6}$$

where V is the external potential applied (function of time t and position $\mathbf{r}$) and C is a constant proportional to the number of atoms in the condensate and to the scattering length. Equation (2.5) is to be solved to describe the condensate evolution from a given initial condensate state.

2.2. **The Interaction Picture method.** The Interaction Picture (IP) method for solving the PDE (2.1) under appropriate initial condition may be understood as follows. The interval $I =]0, S[$ is divided into K sub-intervals where the grid points are denoted s_k, $k = \{0, \ldots, K\}$ such that $]0, S] = \cup_{k=0}^{K-1}]s_k, s_{k+1}]$ where $0 = s_0 < s_1 < \cdots < s_{K-1} < s_K = S$. For all $k \in \{0, \ldots, K-1\}$ the step length between s_k and s_{k+1} is denoted h_k and we also set $s_{k+\frac{1}{2}} = s_k + \frac{h_k}{2}$.

Solving equation (2.1) for the initial condition $u(s = 0, r) = \nu_0(r)$, $\forall r \in \Omega$, is equivalent to solving the following sequence of connected problems:

$$\left\{ \begin{array}{l} \dfrac{\partial}{\partial s} u_0(s,r) = \mathcal{D}\, u_0(s,r) + \mathcal{N}(u_0)(s,r) \quad \forall s \in]s_0, s_1]\ \forall r \in \Omega \\ \quad u_0(s_0, r) = \nu_0(r) \quad \forall r \in \Omega \end{array} \right. \tag{2.7}$$

and $\forall k \in \{1, \ldots, K-1\}$

$$\left\{ \begin{array}{l} \dfrac{\partial}{\partial s} u_k(s,r) = \mathcal{D}\, u_k(s,r) + \mathcal{N}(u_k)(s,r) \quad \forall s \in]s_k, s_{k+1}]\ \forall r \in \Omega \\ \quad u_k(s_k, r) = u_{k-1}(s_k, r) \quad \forall r \in \Omega \end{array} \right. \tag{2.8}$$

Obviously for all $k \in \{0, \ldots, K-1\}$ the unknown functions u and u_k are related by

$$\forall s \in [s_k, s_{k+1}] \quad \forall r \in \Omega \quad u(s,r) = u_k(s,r).$$

Let us consider one of the problems defined in (2.7)–(2.8) for a given value of $k \in \{0, \ldots, K-1\}$. Such a problem reads

$$\left\{ \begin{array}{l} \dfrac{\partial}{\partial s} u_k(s,r) = \mathcal{D}\, u_k(s,r) + \mathcal{N}(u_k)(s,r) \quad \forall s \in]s_k, s_{k+1}]\ \forall r \in \Omega \\ \quad u_k(s_k, r) = \nu_k(r) \quad \forall r \in \Omega \end{array} \right. \tag{2.9}$$

where ν_k is a given function. We introduce as new unknown the mapping

$$u_k^{\mathrm{ip}} : (s,r) \in [s_k, s_{k+1}] \times \Omega \longmapsto \mathrm{e}^{-(s - s_{k+\frac{1}{2}})\mathcal{D}} \cdot u_k(s,r), \tag{2.10}$$

where the exponential term has to be understood in the sense of the continuous group generated by the unbounded linear operator $\mathcal{D}$ [19, 20]. From (2.9) one can show [20] that the new unknown u_k^{ip} is the solution to the following problem

$$\left\{ \begin{array}{l} \dfrac{\partial}{\partial s} u_k^{\mathrm{ip}}(s,r) = \mathcal{G}_k(s, r, u_k^{\mathrm{ip}}(s,r)) \quad \forall s \in]s_k, s_{k+1}]\ \forall r \in \Omega \\ \quad u_k^{\mathrm{ip}}(s_k, r) = \mathrm{e}^{\frac{h_k}{2}\mathcal{D}} \cdot \nu_k(r) \quad \forall r \in \Omega \end{array} \right. \tag{2.11}$$

where $\mathcal{G}_k(s, r, \cdot) = \mathrm{e}^{-(s - s_{k+\frac{1}{2}})\mathcal{D}} \circ \mathcal{N} \circ \mathrm{e}^{(s - s_{k+\frac{1}{2}})\mathcal{D}}$. The major interest for using the change of unknown (2.10) is that on the contrary to problem (2.9), problem (2.11) for the unknown u_k^{ip} does not anymore involve explicitly partial derivation with respect to the variable r. Partial derivation with respect to the variable r now occurs through the operator $\mathrm{e}^{\pm(s - s_{k+\frac{1}{2}})\mathcal{D}}$ which is computed separately. Thus problem (2.11) can be numerically solved just as if it was a

nonlinear ODE with r as a parameter using a standard quadrature scheme for ODE such as Runge-Kutta (RK) schemes.

When solving problem (2.11) a first stage consists in computing the initial condition data function $u_k^{\mathrm{ip}}(s_k,\cdot) : r \mapsto \mathrm{e}^{\frac{h_k}{2}\mathcal{D}} \cdot \nu_k(r)$. This can be done by solving the following linear PDE problem [20]

$$\left\{ \begin{array}{l} \dfrac{\partial}{\partial s} v_k(s,r) = \mathcal{D}\, v_k(s,r) \qquad \forall s \in]s_k, s_{k+\frac{1}{2}}] \; \forall r \in \Omega \\ \quad v_k(s_k,r) = \nu_k(r) \qquad \forall r \in \Omega \end{array} \right. \tag{2.12}$$

since we have $u_k^{\mathrm{ip}}(s_k,\cdot) = v_k(s_{k+\frac{1}{2}},\cdot)$. Once the solution to problem (2.11) has been computed, the inverse mapping of (2.10) has to be used to get the solution to problem (2.9) at grid point s_{k+1}. The mapping $u_k(s_{k+1},\cdot) : r \mapsto \mathrm{e}^{-\frac{h_k}{2}\mathcal{D}} \cdot u_k^{\mathrm{ip}}(r)$ coincides with the solution at grid point s_{k+1} to the following linear PDE problem

$$\left\{ \begin{array}{l} \dfrac{\partial}{\partial s} w_k(s,r) = \mathcal{D}\, w_k(s,r) \qquad \forall s \in]s_{k+\frac{1}{2}}, s_{k+1}] \; \forall r \in \Omega \\ \quad w_k(s_k,r) = u_k^{\mathrm{ip}}(s_{k+1},r) \qquad \forall r \in \Omega \end{array} \right. \tag{2.13}$$

In the same way, the mappings $s \mapsto \mathrm{e}^{-(s-s_{k+\frac{1}{2}})\mathcal{D}}$ and $s \mapsto \mathrm{e}^{(s-s_{k+\frac{1}{2}})\mathcal{D}}$ involved in the definition of the operator $\mathcal{G}_k$ can be evaluated at any intermediate computational grid points by solving linear PDE problems analogous to (2.12) and (2.13). Thus for each step k a major part of the computational effort lies in the resolution of the linear PDE problems (2.12) and (2.13). The numerical method used to solve them is strongly dependent on the linear operator $\mathcal{D}$ and domain Ω, that is to say to the physical application under consideration. For the GPE, this PDE problem is a heat type problem set in a 2D or 3D domain. In [1, 2, 18] it is solved by a Fourier spectral method. For the GNLSE, problems (2.12) and (2.13) where $\Omega = \mathbb{R}$ can be solved by a direct use of Fourier transforms [8]. As well the cost of the evaluation of the terms involving the nonlinear operator $\mathcal{N}$ is strongly dependent on the physical application. Nevertheless it is a direct function evaluation without intermediate PDE problem to be solved. Thus, in designing a new embedded RK method for adaptive step-size control purposes in the IP method we have to keep in mind that the global computational cost of the method will be directly proportional to the number of exponential operators and to the number of nonlinear operators $\mathcal{N}$ involved in the numerical scheme.

To conclude this section we may compare the IP method approach to the Symmetric Split-step (SS) one which is based on the use of Strang splitting formula [21]. Using the same framework as the one presented above for the IP method, the Symmetric Split-step method consists in solving over each sub-interval $[s_k, s_{k+1}]$ for $k \in \{0,\ldots,K-1\}$ the following 3 nested problems:

$$\left\{ \begin{array}{l} \dfrac{\partial}{\partial s} v_k(s,r) = \mathcal{D}\, v_k(s,r) \qquad \forall s \in]s_k, s_{k+\frac{1}{2}}] \; \forall r \in \Omega \\ \quad v_k(s_k,r) = u_{k-1}(s_k,r) \qquad \forall r \in \Omega \end{array} \right. \tag{2.14}$$

where $u_{k-1}(s_k,\cdot) : r \mapsto u_{k-1}(s_k,r)$ represents the solution to problem (2.9) at grid point s_k computed by the numerical scheme at the previous step $k-1$; then

$$\left\{\begin{array}{ll} \dfrac{\partial}{\partial s}u_k^{\mathrm{SS}}(s,r) = \mathcal{N}(u_k^{\mathrm{SS}})(s,r) & \forall s \in]s_k, s_{k+1}] \; \forall r \in \Omega \\ u_k^{\mathrm{SS}}(s_k,r) = v_k(s_{k+\frac{1}{2}},r) & \forall r \in \Omega \end{array}\right. \tag{2.15}$$

where $v_k(s_{k+\frac{1}{2}},\cdot) : r \mapsto v_k(s_{k+\frac{1}{2}},r)$ represents the solution to problem (2.14) at half grid point $s_{k+\frac{1}{2}}$; and finally

$$\left\{\begin{array}{ll} \dfrac{\partial}{\partial s}w_k(s,r) = \mathcal{D}\, w_k(s,r) & \forall s \in]s_{k+\frac{1}{2}}, s_{k+1}] \; \forall r \in \Omega \\ w_k(s_{k+\frac{1}{2}},r) = u_k^{\mathrm{SS}}(s_{k+1},r) & \forall r \in \Omega \end{array}\right. \tag{2.16}$$

where $u_k^{\mathrm{SS}}(s_{k+1},\cdot) : r \mapsto u_k^{\mathrm{SS}}(s_{k+1},r)$ represents the solution to problem (2.15) at node s_{k+1}. The approximate solution to problem (2.9) at grid point s_{k+1} is given by $u_k(s_{k+1},\cdot) \approx w_k(s_{k+1},\cdot)$. We can observe that problem (2.12) in the IP method coincides with problem (2.14) in the SS method whereas problem (2.13) in the IP method coincides with problem (2.16) in the SS method. Moreover problem (2.11) in the IP method and problem (2.15) in the SS method only differ by the function involved in the right hand side of the ordinary differential equation. However whereas the splitting approach involved in the IP method is exact since it corresponds to a change of unknown, the accuracy of the SS method is dependent on the second order convergence of the Strang splitting formula [21].

3. Embedded Runge-Kutta 5(4) formulae for the IP method

3.1. **Overview of embedded RK5(4) schemes.** We recall that it has been proved (see e.g. [22]) that it does not exist 5th order RK formula with only 5 computational stages and more generally that for $p \geq 5$ no explicit RK method exists of order p with $s = p$ stages (see e.g. [10] thm. 5.5 for a proof of the statement). The minimum number of stages for a 5th order RK formula is 6. The coefficients of a s stages RK formula can be described in a very concise manner in an array (termed a *Butcher tableau*) in the form [9]

$$\begin{array}{c|c} c & \mathrm{A} \\ \hline & b^\top \end{array} \tag{3.1}$$

where A is a s by s matrix with real entries and b and c are 2 real vectors in $\mathbb{R}^s$. The real numbers b_i, $i = 1,\ldots,s$ are termed the *weight* of the RK scheme whereas the real numbers c_i, $i = 1,\ldots,s$ correspond to the *elementary quadrature nodes* of the RK scheme. For a RK formula to be of order p, conditions must be satisfied by the entries of matrix A and vectors b and c. The number of conditions, termed *order conditions* in the sequel, to be satisfied for the 4th order is 8 whereas for the 5th order this number raises up to 17.

Embedded Runge-Kutta (ERK) schemes are special RK schemes designed to deliver two approximations of the solution of the ODE under consideration, corresponding to 2 RK schemes of different convergence orders p and q ($q > p$ and most of the time $q = p+1$). These 2

approximations of the solution can be considered as an accurate approximate solution (the one computed with the numerical scheme of higher order q) and a coarse approximate solution (the one computed with the one of lower order p). These 2 approximate solutions obtained with RK schemes of different orders can be combined in a specific way so as to deliver an estimation of the local error committed while approaching the solution with the lower order method [9, 10, 11]. In practise even if the local error estimate obtained with the RK pair holds only for the lower order method, the value given by the higher order method is used as the approximation of the solution for the subsequent computations since its accuracy is better than the one obtained from the lower order method. This approach, referred in the literature as the *local extrapolation mode* for ERK methods, slightly overestimates the actual local error [9, 10, 11].

In [23], E. Fehlberg was interested in the construction of RK pairs of order $p = 4$ and $q = 5$ under simplifying assumptions (to reduce the number of order conditions to be taken into account when considering the 5th order formula) and he proposed a very popular ERK scheme now referred in the literature as the *Fehlberg 4(5) formula* [9, 10, 11]. In Fehlberg approach, the lower order approximation was intended to be used as an initial value for the next step, so in order to make his method optimal E. Fehlberg imposed conditions on RK coefficients in order to minimize the 5th *order truncation error coefficients* for the lower order result. This approach has the disadvantage of providing an estimation of the local error substantially smaller than the true one when the local extrapolation mode is used. The first efforts at constructing RK pairs of order $p = 4$ and $q = 5$ (ERK5(4) scheme) that minimize the truncation error coefficients of the higher order RK formula were undertaken by J.R. Dormand and P.J. Prince [24]. Their approach consists in looking for a 4th order RK formula embedded in a 5th order RK formula defined in $s = 7$ stages. Compared to an ERK5(4) scheme with only 6 stages, this approach offers more flexibility for determining the values of the RK coefficients aimed at minimizing the 6th order truncation error coefficients. The computational extra cost of this supplementary stage is counterbalance by using the so-called FSAL (First Step At Last) property. The FSAL property imposes that the vector b corresponding to the output approximation coefficients has its last component 0 and its other components identical to the last row of the matrix A. The consequence is that while the Dormand and Prince ERK5(4) scheme has 7 stages, it operates as though it only has 6 stages because the evaluation of the seventh and last stage can be retained to serve as the first stage of the next step.

To demonstrate the construction of an ERK scheme suited for use in conjunction with the IP method we follow the idea which allowed J.R. Dormand and P.J. Prince to construct their famous family of ERK5(4) schemes. However the method we are constructing is not of Dormand and Prince type since we use different additional conditions for fully determining the coefficients of the RK pair in order to "optimize" our ERK5(4) scheme for a use with the IP method. The construction of our method is now detailed.

3.2. **Conditions for an embedded RK5(4) scheme.** For the higher order RK formula, we consider an explicit 5th order formula with 7 stages given by a Butcher tableau in the form (3.1)

where A is a real lower triangular matrix with 0 diagonal entries

$$\mathrm{A} = \begin{pmatrix} & & & & & \\ a_{2,1} & & & & & \\ a_{3,1} & a_{3,2} & & & & \\ a_{4,1} & a_{4,2} & a_{4,3} & & & \\ a_{5,1} & a_{5,2} & a_{5,3} & a_{5,4} & & \\ a_{6,1} & a_{6,2} & a_{6,3} & a_{6,4} & a_{6,5} & \\ a_{7,1} & a_{7,2} & a_{7,3} & a_{7,4} & a_{7,5} & a_{7,6} \end{pmatrix} \tag{3.2}$$

and $b = (\widehat{b}_1, \ldots, \widehat{b}_7)^\top$, $c = (c_1, \ldots, c_7)^\top$ (where the symbol $^\top$ stands for the transpose) are real vectors, whereas as a lower order RK formula we consider an explicit 4th order formula defined by the same matrix A and vector c but with a different vector $b = (b_1, \ldots, b_7)^\top$. Usually, for the study of high order RK formulae, the following conditions are imposed

$$c_1 = 0 \quad \text{and} \quad \forall i \in \{2, \ldots, 7\} \quad c_i = \sum_{j=1}^{i-1} a_{ij}. \tag{3.3}$$

These assumptions greatly simplify the derivation of order conditions for high order RK formulae although they are not necessary. Assumptions (3.3) express that the function evaluations corresponding to the internal stages of the RK formula, which contribute to the estimation of the endpoint solution, provide a cost free second order approximation at these nodes for the ODE problem (2.11). These assumptions were already imposed by W. Kutta in his seminal work [25].

The order condition equations to be satisfied for the RK formula defined by Butcher tableau (3.1) to be of order 4 read[1]

$$\forall k \in \{1, \ldots, 4\} \quad \forall j \in \{1, \ldots, \theta_k\} \quad a_j^{(k)} = 0, \tag{3.4}$$

where

$$a_1^{(1)} = \sum_{i=1}^{7} b_i - 1 \tag{3.5}$$

$$a_1^{(2)} = \sum_{i=1}^{7} b_i c_i - \frac{1}{2} \tag{3.6}$$

$$a_1^{(3)} = \frac{1}{2} \sum_{i=1}^{7} b_i c_i^2 - \frac{1}{6} \tag{3.7}$$

$$a_2^{(3)} = \sum_{i,j=1}^{7} b_i a_{i,j} c_j - \frac{1}{6} \tag{3.8}$$

$$a_1^{(4)} = \frac{1}{6} \sum_{i=1}^{7} b_i c_i^3 - \frac{1}{24} \tag{3.9}$$

$$a_2^{(4)} = \sum_{i,j=1}^{7} b_i c_i a_{i,j} c_j - \frac{1}{8} \tag{3.10}$$

$$a_3^{(4)} = \frac{1}{2} \sum_{i,j=1}^{7} b_i a_{i,j} c_j^2 - \frac{1}{24} \tag{3.11}$$

$$a_4^{(4)} = \sum_{i,j,k=1}^{7} b_i a_{i,j} a_{j,k} c_k - \frac{1}{24} \tag{3.12}$$

[1]The value of the θ_k are given by thm. 302B of [9]. We have $\theta_1 = \theta_2 = 1$, $\theta_3 = 2$, $\theta_4 = 4$, $\theta_5 = 9$ and $\theta_6 = 20$.

Remark 1. *The order condition equation* (3.5) *reads* $\sum_{i=1}^{7} b_i = 1$ *and it is known to ensure the consistency of the RK method* [9, 10, 11].

The conditions for a RK formula to be of order 5 include the previous order condition equations where b_i is replaced with $\widehat{b}_i$ – these conditions will be denoted by $\widehat{a}_1^{(1)}, \ldots, \widehat{a}_4^{(4)}$ in the sequel and numbered $\widehat{(3.5)}$ to $\widehat{(3.12)}$ – together with the following order condition equations specific to the 5th order

$$\forall j \in \{1, \ldots, \theta_5\} \quad \widehat{a}_j^{(5)} = 0, \tag{3.13}$$

where

$$\widehat{a}_1^{(5)} = \frac{1}{24} \sum_{i=1}^{7} \widehat{b}_i c_i^4 - \frac{1}{120} \tag{3.14}$$

$$\widehat{a}_2^{(5)} = \frac{1}{2} \sum_{i,j=1}^{7} \widehat{b}_i c_i^2 a_{i,j} c_j - \frac{1}{20} \tag{3.15}$$

$$\widehat{a}_3^{(5)} = \frac{1}{2} \sum_{i,j,k=1}^{7} \widehat{b}_i a_{i,j} c_j a_{i,k} c_k - \frac{1}{40} \tag{3.16}$$

$$\widehat{a}_4^{(5)} = \frac{1}{2} \sum_{i,j=1}^{7} \widehat{b}_i c_j^2 a_{i,j} c_i - \frac{1}{30} \tag{3.17}$$

$$\widehat{a}_5^{(5)} = \frac{1}{6} \sum_{i,j=1}^{7} \widehat{b}_i a_{i,j} c_j^3 - \frac{1}{120} \tag{3.18}$$

$$\widehat{a}_6^{(5)} = \sum_{i,j,k=1}^{7} \widehat{b}_i a_{i,j} c_i a_{j,k} c_k - \frac{1}{30} \tag{3.19}$$

$$\widehat{a}_7^{(5)} = \sum_{i,j,k=1}^{7} \widehat{b}_i a_{i,j} c_j a_{j,k} c_k - \frac{1}{40} \tag{3.20}$$

$$\widehat{a}_8^{(5)} = \frac{1}{2} \sum_{i,j,k=1}^{7} \widehat{b}_i a_{i,j} a_{j,k} c_k^2 - \frac{1}{120} \tag{3.21}$$

$$\widehat{a}_9^{(5)} = \sum_{i,j,k,m=1}^{7} \widehat{b}_i a_{i,j} a_{j,k} a_{k,m} c_m - \frac{1}{20} \tag{3.22}$$

The truncation error coefficients of a 4th order RK formula is defined as $\|a^{(5)}\|_2 = \left(\sum_{j=1}^{9} (a_j^{(5)})^2\right)^{\frac{1}{2}}$ where $a_j^{(5)}, j = 1, \ldots, 9$ are given by relations (3.14) to (3.22) with $\widehat{b}_i$ replaced by b_i. The truncation error coefficients of a 5th order RK formula will be defined in Section 3.4.1.

3.3. **Conditions related to the efficiency of the IP method.** In the Interaction Picture method, one step of the 5th order RK method is used to approach the solution to problem (2.11) as follows:

$$\forall r \in \Omega \qquad u_k^{\mathrm{ip}}(s_{k+1}, r) \approx \widetilde{u}_{k+1}^{\mathrm{ip},(5)}(r),$$

where

$$\widetilde{u}_{k+1}^{\mathrm{ip},(5)}(r) = u_k^{\mathrm{ip}}(s_k, r) + h_k \left(\widehat{b}_1\, \alpha_1 + \widehat{b}_2\, \alpha_2 + \widehat{b}_3\, \alpha_3 + \widehat{b}_4\, \alpha_4 + \widehat{b}_5\, \alpha_5 + \widehat{b}_6\, \alpha_6 + \widehat{b}_7\, \alpha_7 \right) \tag{3.23}$$

and

$$\begin{aligned} \alpha_1 &= \mathcal{G}_k(s_k, r, u_k^{\mathrm{ip}}(s_k, r)) = \mathrm{e}^{\frac{h_k}{2}\mathcal{D}} \cdot \mathcal{N}(\mathrm{e}^{-\frac{h_k}{2}\mathcal{D}} \cdot u_k^{\mathrm{ip}}(s_k, r)) \\ &= \mathrm{e}^{\frac{h_k}{2}\mathcal{D}} \cdot \mathcal{N}(u_k(s_k, r)) \end{aligned}$$

$$
\begin{aligned}
\alpha_2 &= \mathcal{G}_k(s_k + c_2 h_k, r, u_k^{\mathrm{ip}}(s_k, r) + h_k \alpha_1 a_{2,1}) \\
&= \mathrm{e}^{-(c_2-\frac{1}{2})h_k\mathcal{D}} \cdot \mathcal{N}\left(\mathrm{e}^{(c_2-\frac{1}{2})h_k\mathcal{D}} \cdot \left[u_k^{\mathrm{ip}}(s_k, r) + h_k \alpha_1 a_{2,1}\right]\right) \\
\alpha_3 &= \mathcal{G}_k(s_k + c_3 h_k, r, u_k^{\mathrm{ip}}(s_k, r) + h_k(\alpha_1 a_{3,1} + \alpha_2 a_{3,2})) \\
&= \mathrm{e}^{-(c_3-\frac{1}{2})h_k\mathcal{D}} \cdot \mathcal{N}\Big(\mathrm{e}^{(c_3-\frac{1}{2})h_k\mathcal{D}} \cdot \left[u_k^{\mathrm{ip}}(s_k, r) + h_k \textstyle\sum_{j=1}^{2} \alpha_j a_{3,j})\right]\Big) \\
\alpha_4 &= \mathcal{G}_k(s_k + c_4 h_k, r, u_k^{\mathrm{ip}}(s_k, r) + h_k \textstyle\sum_{j=1}^{3} \alpha_j a_{4,j}) \\
&= \mathrm{e}^{-(c_4-\frac{1}{2})h_k\mathcal{D}} \cdot \mathcal{N}\Big(\mathrm{e}^{(c_4-\frac{1}{2})h_k\mathcal{D}} \cdot \left[u_k^{\mathrm{ip}}(s_k, r) + h_k \textstyle\sum_{j=1}^{3} \alpha_j a_{4,j}\right]\Big) \\
\alpha_5 &= \mathcal{G}_k(s_k + c_5 h_k, r, u_k^{\mathrm{ip}}(s_k, r) + h_k \textstyle\sum_{j=1}^{4} \alpha_j a_{5,j}) \\
&= \mathrm{e}^{-(c_5-\frac{1}{2})h_k\mathcal{D}} \cdot \mathcal{N}\Big(\mathrm{e}^{(c_5-\frac{1}{2})h_k\mathcal{D}} \cdot \left[u_k^{\mathrm{ip}}(s_k, r) + h_k \textstyle\sum_{j=1}^{4} \alpha_j a_{5,j}\right]\Big) \\
\alpha_6 &= \mathcal{G}_k(s_k + c_6 h_k, r, u_k^{\mathrm{ip}}(s_k, r) + h_k \textstyle\sum_{j=1}^{5} \alpha_j a_{6,j}) \\
&= \mathrm{e}^{-(c_6-\frac{1}{2})h_k\mathcal{D}} \cdot \mathcal{N}\Big(\mathrm{e}^{(c_6-\frac{1}{2})h_k\mathcal{D}} \cdot \left[u_k^{\mathrm{ip}}(s_k, r) + h_k \textstyle\sum_{j=1}^{5} \alpha_j a_{6,j}\right]\Big) \\
\alpha_7 &= \mathcal{G}_k(s_k + c_7 h_k, r, u_k^{\mathrm{ip}}(s_k, r) + h_k \textstyle\sum_{j=1}^{5} \alpha_j a_{7,j}) \\
&= \mathrm{e}^{-(c_7-\frac{1}{2})h_k\mathcal{D}} \cdot \mathcal{N}\Big(\mathrm{e}^{(c_7-\frac{1}{2})h_k\mathcal{D}} \cdot \left[u_k^{\mathrm{ip}}(s_k, r) + h_k \textstyle\sum_{j=1}^{6} \alpha_j a_{7,j}\right]\Big)
\end{aligned}
$$

Moreover, a 4th order approximate solution to problem (2.11) at grid point s_{k+1} is given by

$$
\widetilde{u}_{k+1}^{\mathrm{ip},(4)}(r) = u_k^{\mathrm{ip}}(s_k, r) + h_k \left(b_1\,\alpha_1 + b_2\,\alpha_2 + b_3\,\alpha_3 + b_4\,\alpha_4 + b_5\,\alpha_5 + b_6\,\alpha_6 + b_7\,\alpha_7\right). \quad (3.24)
$$

By using the change of unknown (2.10) we deduce that the mapping $r \mapsto u_k(s_{k+1}, r)$ solution to problem (2.1) at grid point s_{k+1} can be approximated using the 5th order formula, $\forall r \in \Omega$, by

$$
\widetilde{u}_{k+1}^{(5)}(r) = \mathrm{e}^{\frac{h_k}{2}\mathcal{D}} \cdot \widetilde{u}_{k+1}^{\mathrm{ip},(5)}(r) = \mathrm{e}^{\frac{h_k}{2}\mathcal{D}} \cdot \left(u_k^{\mathrm{ip}}(s_k, r) + h_k\, \widehat{\Phi}(s_k, u_k^{\mathrm{ip}}; h_k)\right), \quad (3.25)
$$

where

$$
\widehat{\Phi}(s_k, u_k^{\mathrm{ip}}; h_k) = \widehat{b}_1\,\alpha_1 + \widehat{b}_2\,\alpha_2 + \widehat{b}_3\,\alpha_3 + \widehat{b}_4\,\alpha_4 + \widehat{b}_5\,\alpha_5 + \widehat{b}_6\,\alpha_6 + \widehat{b}_7\,\alpha_7 \quad (3.26)
$$

is the incremental function of the RK formula. Actually we are only interested in computing an approximate solution for problem (2.1) and the use of the new unknown u_k^{ip} and its approximations $\widetilde{u}_{k+1}^{\mathrm{ip},(5)}$ and $\widetilde{u}_{k+1}^{\mathrm{ip},(4)}$ is a go-between in the computational approach. We can therefore recast the above computational procedure as follows to reduce the computational cost of the method. At step k, the 2 approximate solutions $\widetilde{u}_{k+1}^{(4)}$ and $\widetilde{u}_{k+1}^{(5)}$ are obtained by computing successively

$$
\begin{aligned}
\widetilde{u}_k^{\mathrm{ip},(5)}(r) &= \mathrm{e}^{\frac{h_k}{2}\mathcal{D}} \cdot \widetilde{u}_k^{(5)}(r) \\
\alpha_1 &= \mathrm{e}^{\frac{h_k}{2}\mathcal{D}} \cdot \mathcal{N}(\widetilde{u}_k^{(5)}(r)) \\
\alpha_2 &= \mathrm{e}^{-(c_2-\frac{1}{2})h_k\mathcal{D}} \cdot \mathcal{N}\left(\mathrm{e}^{(c_2-\frac{1}{2})h_k\mathcal{D}} \cdot \left[\widetilde{u}_k^{\mathrm{ip},(5)}(r) + h_k \alpha_1 a_{2,1}\right]\right)
\end{aligned}
$$

$$\alpha_3 = \mathrm{e}^{-(c_3-\frac{1}{2})h_k\mathcal{D}}\cdot\mathcal{N}\Big(\mathrm{e}^{(c_3-\frac{1}{2})h_k\mathcal{D}}\cdot\big[\widetilde{u}_k^{\mathrm{ip},(5)}(r)+h_k\textstyle\sum_{j=1}^{2}\alpha_j a_{3,j}\big)\big]\Big)$$
$$\alpha_4 = \mathrm{e}^{-(c_4-\frac{1}{2})h_k\mathcal{D}}\cdot\mathcal{N}\Big(\mathrm{e}^{(c_4-\frac{1}{2})h_k\mathcal{D}}\cdot\big[\widetilde{u}_k^{\mathrm{ip},(5)}(r)+h_k\textstyle\sum_{j=1}^{3}\alpha_j a_{4,j}\big]\Big)$$
$$\alpha_5 = \mathrm{e}^{-(c_5-\frac{1}{2})h_k\mathcal{D}}\cdot\mathcal{N}\Big(\mathrm{e}^{(c_5-\frac{1}{2})h_k\mathcal{D}}\cdot\big[\widetilde{u}_k^{\mathrm{ip},(5)}(r)+h_k\textstyle\sum_{j=1}^{4}\alpha_j a_{5,j}\big]\Big)$$
$$\alpha_6 = \mathrm{e}^{-(c_6-\frac{1}{2})h_k\mathcal{D}}\cdot\mathcal{N}\Big(\mathrm{e}^{(c_6-\frac{1}{2})h_k\mathcal{D}}\cdot\big[\widetilde{u}_k^{\mathrm{ip},(5)}(r)+h_k\textstyle\sum_{j=1}^{5}\alpha_j a_{6,j}\big]\Big)$$
$$\alpha_7 = \mathrm{e}^{-(c_7-\frac{1}{2})h_k\mathcal{D}}\cdot\mathcal{N}\Big(\mathrm{e}^{(c_7-\frac{1}{2})h_k\mathcal{D}}\cdot\big[\widetilde{u}_k^{\mathrm{ip},(5)}(r)+h_k\textstyle\sum_{j=1}^{6}\alpha_j a_{7,j}\big]\Big).$$

The 4th order approximate solution at grid point s_{k+1} is then given $\forall r\in\Omega$ by

$$\widetilde{u}_{k+1}^{(4)}(r) = \mathrm{e}^{\frac{h_k}{2}\mathcal{D}}\cdot\big(\widetilde{u}_k^{\mathrm{ip},(5)}(r)+h_k(b_1\,\alpha_1+b_2\,\alpha_2+b_3\,\alpha_3+b_4\,\alpha_4 + b_5\,\alpha_5+b_6\,\alpha_6+b_7\,\alpha_7)\big) \tag{3.27}$$

whereas the 5th order approximate solution is given $\forall r\in\Omega$ by

$$\widetilde{u}_{k+1}^{(5)}(r) = \mathrm{e}^{\frac{h_k}{2}\mathcal{D}}\cdot\big(\widetilde{u}_k^{\mathrm{ip},(5)}(r)+h_k(\widehat{b}_1\,\alpha_1+\widehat{b}_2\,\alpha_2+\widehat{b}_3\,\alpha_3+\widehat{b}_4\,\alpha_4 + \widehat{b}_5\,\alpha_5+\widehat{b}_6\,\alpha_6+\widehat{b}_7\,\alpha_7)\big). \tag{3.28}$$

In order to reduce the computational effort, we impose the following values

$$\widehat{b}_7 = 0, \qquad c_7 = 1, \qquad \forall j\in\{1,\ldots,6\}\quad a_{7,j}=\widehat{b}_j. \tag{3.29}$$

With this choice, the function where the nonlinear operator $\mathcal{N}$ acts in the expression of α_7 coincides with $\widetilde{u}_{k+1}^{(5)}$ given by (3.28) and need only to be evaluated one time. Moreover, the value of α_7 computed at step k coincides with the value of α_1 for the next step $k+1$ which save one evaluation of function $\mathcal{G}_k$. Actually, this saving is effective only when the current step is not rejected which is likely to occur the most frequently. Therefore although the ERK5(4) scheme is a 7 stage method, it has the computational cost of a 6 stage method when the current step is accepted. Condition (3.29) is referred in the literature as the FSAL (First Step At Last) property [9, 10].

Under the FSAL assumption (3.29), Butcher tableau for the ERK4(5) scheme corresponds to a matrix A and a vector c in the form

$$c = \begin{pmatrix} 0\\ c_2\\ c_3\\ c_4\\ c_5\\ c_6\\ 1 \end{pmatrix} \qquad \mathrm{A} = \begin{pmatrix} & & & & & \\ a_{2,1} & & & & & \\ a_{3,1} & a_{3,2} & & & & \\ a_{4,1} & a_{4,2} & a_{4,3} & & & \\ a_{5,1} & a_{5,2} & a_{5,3} & a_{5,4} & & \\ a_{6,1} & a_{6,2} & a_{6,3} & a_{6,4} & a_{6,5} & \\ \widehat{b}_1 & \widehat{b}_2 & \widehat{b}_3 & \widehat{b}_4 & \widehat{b}_5 & \widehat{b}_6 \end{pmatrix} \tag{3.30}$$

and for the 5th RK formula we have $b=(\widehat{b}_1,\widehat{b}_2,\widehat{b}_3,\widehat{b}_4,\widehat{b}_5,\widehat{b}_6,0)^\top$ whereas for the 4th order RK formula we have $b=(b_1,b_2,b_3,b_4,b_5,b_6,b_7)^\top$. It is usual for such an ERK scheme to give

the coefficients of the 2 RK formulae in a unique array (termed an *extended Butcher tableau*) as follows

$$\begin{array}{c|ccccccc}
0 & & & & & & & \\
c_2 & a_{2,1} & & & & & & \\
c_3 & a_{3,1} & a_{3,2} & & & & & \\
c_4 & a_{4,1} & a_{4,2} & a_{4,3} & & & & \\
c_5 & a_{5,1} & a_{5,2} & a_{5,3} & a_{5,4} & & & \\
c_6 & a_{6,1} & a_{6,2} & a_{6,3} & a_{6,4} & a_{6,5} & & \\
\hline
1 & \widehat{b}_1 & \widehat{b}_2 & \widehat{b}_3 & \widehat{b}_4 & \widehat{b}_5 & \widehat{b}_6 & \\
 & b_1 & b_2 & b_3 & b_4 & b_5 & b_6 & b_7
\end{array} \tag{3.31}$$

where the gray cells correspond to Butcher tableau for the 5th order RK formula and the whole array is Butcher tableau for the 4th order RK formula.

In order to reduce further the computations, it would be interesting to have several coefficients c_i, $i = 2, \ldots, 6$ equal since the number of exponential terms $\mathrm{e}^{\pm(c_i - \frac{1}{2})h_k\mathcal{D}}$ to evaluate in the ERK5(4)-IP method would be lower. However this feature can not be exploited at this stage and we will now investigate the condition equations for the RK pair defined by Butcher tableau (3.31) to be of order 5 and 4 respectively.

3.4. **Solving the order condition equations.** To ensure that the approximations (3.27) and (3.28) correspond to respectively a 4th order and a 5th order RK formula, the values of the remaining free coefficients have to be chosen in order that, together with conditions (3.3), the 8 order condition equations (3.5)–(3.12) hold for the 4th order RK formula and that the 17 order condition equations $\widehat{(3.5)}$–$\widehat{(3.22)}$ hold for the 5th order RK formula. This is a very tedious task to solve this system of 31 nonlinear equations even with the help of symbolic calculus softwares. In [26] a complete characterisation of the 17 order conditions for 5th order RK formulae is given; however the results in this study have not found any practical implementation. Fortunately, the system of order condition equations corresponds to a system of necessary and sufficient conditions for a 5th order formula and it can be replaced by a much simpler system of sufficient conditions. Thus, a solution of the 17 order condition equations for the 5th order RK method is considered by imposition of the following additional conditions [22, 24]

$$\widehat{b}_2 = 0, \tag{3.32}$$

$$\forall j \in \{1, \ldots, 7\} \quad \sum_{i=1}^{7} \widehat{b}_i a_{i,j} = \widehat{b}_j (1 - c_j), \tag{3.33}$$

$$\forall i \in \{3, \ldots, 7\} \quad \sum_{j=1}^{7} c_j a_{i,j} = \frac{1}{2} c_i^2. \tag{3.34}$$

Actually from the choice made in (3.29), condition (3.33) reads

$$\forall j \in \{1, \ldots, 6\} \quad \sum_{i=1}^{6} \widehat{b}_i a_{i,j} = \widehat{b}_j (1 - c_j) \tag{3.35}$$

whereas condition (3.34) can be decomposed in

$$\forall i \in \{3, \ldots, 6\} \quad \sum_{j=1}^{6} c_j a_{i,j} = \frac{1}{2} c_i^2 \tag{3.36}$$

and in order condition equation $\widehat{(3.6)}$. Making $j = 6$ in condition (3.35) we obtain $\widehat{b}_6(1 - c_6) = 0$. We assume that $c_6 = 1$ and $\widehat{b}_6 \neq 0$ otherwise the RK scheme would have 5 stages and it is known that it does not exist 5th order RK methods with only 5 stages [22].

Remark 2. *Conditions* (3.35) *and* (3.36) *correspond to 2 of the 3 additional conditions considered by J.R. Dormand and P.J. Prince in* [24] *in their quest for a class of ERK5(4) schemes. Thus compare to Dormand and Prince work, our approach is more general in the sense that less additional conditions are imposed. This is likely to provide more flexibility in the design of an embedded Runge-Kutta scheme suited for the IP method.*

Considering the additional conditions (3.32)–(3.34), the order condition equations for the 5th order RK formula that still remain to be solved are equations $\widehat{(3.5)}$, $\widehat{(3.6)}$, $\widehat{(3.7)}$, $\widehat{(3.9)}$, $\widehat{(3.14)}$, $\widehat{(3.17)}$, $\widehat{(3.19)}$, where caps are used to indicate conditions for the 5th order RK formula involving the $\widehat{b}_i$, see on p. 247. From condition (3.34), one can show that when order condition $\widehat{(3.17)}$ is satisfied then order condition $\widehat{(3.19)}$ is equivalent to

$$\sum_{i=1}^{6} \widehat{b}_i c_i a_{i,2} = 0. \tag{3.37}$$

We now consider the 8 order condition equations for the 4th order RK formula. Using condition (3.34) and order condition (3.7) one can show that order condition equation (3.8) is satisfied if and only if $b_2\, c_2^2 = 0$. We assume that $c_2 \neq 0$ (for otherwise we would in effect be searching for a RK formula with only 6 quadrature nodes) and therefore we must have $b_2 = 0$. From condition (3.34), one can show that order condition equations (3.9) and (3.10) are equivalent. Moreover condition (3.34) implies that when order condition equation (3.11) is satisfied then order condition equation (3.12) is equivalent to

$$\sum_{i=1}^{6} b_i a_{i,2} = 0. \tag{3.38}$$

We recall that according to order condition $\widehat{(3.7)}$

$$\forall i \in \{2, \ldots, 7\} \qquad c_i = \sum_{j=1}^{i-1} a_{i,j}. \tag{3.39}$$

To summarize, we have set $\widehat{b}_2 = 0$, $\widehat{b}_7 = 0$, $c_7 = 1$, we have obtained that $b_2 = 0$, $c_6 = 1$ and we have to solve the following set of equations : (3.35), (3.36), $\widehat{(3.5)}$, $\widehat{(3.6)}$, $\widehat{(3.7)}$, $\widehat{(3.9)}$, $\widehat{(3.14)}$, $\widehat{(3.17)}$, (3.37), (3.5), (3.6), (3.7), (3.9), (3.11), (3.38) and (3.39) to determine $\widehat{b}_1$, $\widehat{b}_3$, $\widehat{b}_4$, $\widehat{b}_5$ and $\widehat{b}_6$, b_1, b_3, b_4, b_5, b_6 and b_7, c_2, c_3, c_4 and c_5 and $a_{i,j}$ for $1 \leq i < j \leq 6$. Altogether

we have 33 nonlinear equations and 30 unknowns. We have proceeded to the resolution of the nonlinear system with the help of the symbolic calculus software Maple [27]. The values of $\widehat{b}_1$ and b_1 are determined from $\widehat{(3.5)}$ and (3.5) since these are the only equations in which they occur. When c_3, c_4 and c_5 are distinct and not equal to 1, order condition equations $\widehat{(3.6)}$, $\widehat{(3.7)}$, $\widehat{(3.9)}$, $\widehat{(3.14)}$ yield the following expressions for $\widehat{b}_3$, $\widehat{b}_4$, $\widehat{b}_5$ and $\widehat{b}_6$ as function of the parameters c_3, c_4 and c_5

$$\begin{aligned}
\widehat{b}_3 &= -\frac{1}{60}\,\frac{10\,c_4c_5 - 5\,(c_4+c_5)+3}{c_3\,(c_3-1)\,(c_3-c_5)\,(c_3-c_4)},\\
\widehat{b}_4 &= -\frac{1}{60}\,\frac{10\,c_3c_5 - 5\,(c_3+c_5)+3}{c_4\,(c_4-1)\,(c_4-c_3)\,(c_4-c_5)},\\
\widehat{b}_5 &= -\frac{1}{60}\,\frac{10\,c_3c_4 - 5\,(c_3+c_4)+3}{c_5\,(c_5-1)\,(c_5-c_3)\,(c_5-c_4)},\\
\widehat{b}_6 &= \frac{1}{60}\,\frac{30c_3c_4c_5 - 20(c_3c_4+c_3c_5+c_4c_5)+15(c_3+c_4+c_5)-12}{(c_5-1)\,(c_4-1)\,(c_3-1)}.
\end{aligned} \tag{3.40}$$

Now one can see that under order condition equation $\widehat{(3.6)}$, the relations given by condition (3.35) for $j=1$ and $j=3$ are equivalent.

Some of the entries of matrix A can be readily expressed. The coefficient $a_{3,2}$ is determined from (3.36) for $j=3$ and reads $a_{3,2} = c_3^2/2c_2$. Then, condition (3.39) for $j=2$ and $j=3$ gives

$$a_{2,1} = c_2, \qquad a_{3,1} = \frac{c_3(2c_2-c_3)}{2c_2},$$

whereas condition (3.35) for $j=5$ gives

$$a_{6,5} = \frac{\widehat{b}_5(1-c_5)}{\widehat{b}_6}.$$

We also have some simple relations between some of the matrix entries. Condition (3.35) for $j=4$ gives

$$a_{6,4} = \frac{\widehat{b}_4(1-c_4) - \widehat{b}_5 a_{5,4}}{\widehat{b}_6}$$

and condition (3.39) for $j=4$, $j=5$ and $j=6$ gives

$$\begin{aligned}
a_{4,1} &= \frac{2a_{4,2}(c_2-c_3) - c_4^2 + 2\,c_3c_4}{2c_3},\\
a_{5,1} &= c_5 - a_{5,2} - a_{5,3} - a_{5,4},\\
a_{6,1} &= -\frac{\widehat{b}_4(1-c_4) + \widehat{b}_5(1-c_5-a_{5,4}) - \widehat{b}_6(1-a_{6,2}-a_{6,3})}{\widehat{b}_6}.
\end{aligned}$$

We then solve (3.37) and (3.35) for $j = 2$ and we obtain

$$a_{6,2} = \frac{\widehat{b}_3 c_3^2(c_4 - c_3) + 2\widehat{b}_5 c_2(c_4 - c_5)a_{5,2}}{2\widehat{b}_6 c_2\,(1 - c_4)},$$
$$a_{4,2} = \frac{\widehat{b}_3 c_3^2(c_3 - 1) + 2\widehat{b}_5 c_2(c_5 - 1)a_{5,2}}{2\widehat{b}_4 c_2\,(1 - c_4)}.$$

At this stage, order conditions $\widehat{(3.17)}$ and (3.11) as well as conditions (3.35) for $j = 3$, (3.36) for $j = 5$ and $j = 6$ have not been used and coefficients $a_{5,2}$, $a_{5,3}$, $a_{5,4}$ and $a_{6,3}$ remain to be determined. We choose to express $a_{5,4}$ and $a_{6,3}$ in terms of $a_{5,2}$ and $a_{5,3}$ from order conditions (3.36) for $j = 5$ and (3.35) for $j = 3$ respectively. We obtain

$$a_{5,4} = \frac{c_5^2 - 2(c_2 a_{5,2} + c_3 a_{5,3})}{2c_4},$$
$$a_{6,3} = \frac{2\widehat{b}_5 c_2(c_5 - 1)a_{5,2} + \widehat{b}_3 c_3(c_3 - 1)(2c_4 + c_3 - 2) + (c_4 - 1)(\widehat{b}_4 c_4^2 + 2\widehat{b}_5 c_3 a_{5,3})}{2\,(c_4 - 1)\,c_3\widehat{b}_6}.$$

It happens that condition (3.36) for $j = 6$ is now satisfied. Finally, we have to solve order condition equations $\widehat{(3.17)}$ and (3.11) to get $a_{5,2}$ and $a_{5,3}$ in terms of c_2, c_3, c_4 and c_5.

Remark 3. *When considering the special case $c_2 = 1/5$, $c_3 = 3/10$, $c_4 = 4/5$, $c_5 = 8/9$, we obtain Dormand and Price RK5(4)-7M formula* [24].

As mentioned earlier our goal is to obtain embedded 4th order and 5th order RK pair with a maximum number of the c_i having the value $\frac{1}{2}$. We therefore impose $c_2 = c_4 = \frac{1}{2}$ in order to reduce the number of exponential operators involved in the ERK5(4)-IP method, see Section 3.3. One may notice from the expression of the $\widehat{b}_i$ given by (3.40) that in our approach it is not possible to have more than these 2 coefficients equal to $\frac{1}{2}$ since otherwise the denominators would cancelled. Moreover in order to reduce the amount of computation when the embedded Runge-Kutta scheme is used in conjunction with the IP method it is convenient to set $c_3 = \frac{1}{2} - \delta$ and $c_5 = \frac{1}{2} + \delta$ where $\delta \in]0, \frac{1}{2}[$ is a free parameter. Extended Butcher tableau corresponding to our ERK5(4) scheme then reads

$$\begin{array}{c|ccccccc}
0 & & & & & & & \\
\frac{1}{2} & \frac{1}{2} & & & & & & \\
\frac{1}{2} - \delta & (\frac{1}{2} - \delta)(\frac{1}{2} + \delta) & (\frac{1}{2} - \delta)^2 & & & & & \\
\frac{1}{2} & a_{4,1} & a_{4,2} & a_{4,3} & & & & \\
\frac{1}{2} + \delta & a_{5,1} & a_{5,2} & a_{5,3} & a_{5,4} & & & \\
1 & a_{6,1} & a_{6,2} & a_{6,3} & a_{6,4} & a_{6,5} & & \\
\hline
1 & \widehat{b}_1 & 0 & \widehat{b}_3 & \widehat{b}_4 & \widehat{b}_5 & \widehat{b}_6 & \\
\hline
 & b_1 & 0 & b_3 & b_4 & b_5 & b_6 & b_7
\end{array} \tag{3.41}$$

where $a_{5,4} = -\delta\ (1 - 10\,\delta)\,(1 + 2\,\delta)$ and

$$a_{4,1} = \frac{1 - 4\delta - (1+\delta)}{4\,(1-2\delta)^2} \qquad a_{4,2} = \frac{1-6\delta}{4(1-2\delta)}$$

$$a_{4,3} = \frac{\delta}{(1-2\delta)^2} \qquad \widehat{b}_1 = \frac{3 - 20\delta^2}{30(1-4\delta^2)}$$

$$\widehat{b}_3 = \frac{1}{60\delta^2\,(1-4\delta^2)} \qquad \widehat{b}_4 = -\frac{1-20\delta^2}{30\delta^2}$$

$$\widehat{b}_5 = \frac{1}{60\delta^2\,(1-4\delta^2)} \qquad \widehat{b}_6 = \frac{3-20\delta^2}{30(1-4\delta^2)}$$

$$a_{5,1} = \frac{\left(1 - 6\delta + 8\delta^3 + 320\delta^4 - 12\delta^2\right)(1+2\delta)}{4\,(1-2\delta)^2}$$

$$a_{5,2} = \frac{(1+2\delta)\left(1 - 8\delta - 44\delta^2 + 240\delta^3\right)}{4(1-2\delta)}$$

$$a_{5,3} = \frac{(1+2\delta)\,\delta\left(3 - 2\delta - 40\delta^2\right)}{(1-2\delta)^2}$$

$$a_{6,1} = \frac{1 - 12\delta + 4\delta^2 + 80\delta^3 + 160\delta^4}{(-3 + 20\delta^2)\,(1-2\delta)^2}$$

$$a_{6,2} = \frac{2\delta\left(9 - 20\delta - 60\delta^2\right)}{(1-2\delta)\,(3-20\delta^2)}$$

$$a_{6,3} = \frac{1 - 4\delta - 12\delta^2 + 160\delta^4 + 320\delta^5}{4\,(1-2\delta)^2\,\delta^2\,(3-20\delta^2)}$$

$$a_{6,4} = -\frac{1 - \delta - 16\delta^2 + 20\delta^3 + 80\delta^4}{2\delta^2\,(3-20\delta^2)}$$

$$a_{6,5} = \frac{1-2\delta}{4\delta^2\,(3-20\delta^2)}$$

Furthermore, the coefficients b_1, b_3, b_4, b_5 and b_6 are obtained by solving the system of linear equations (3.5), (3.6), (3.7), (3.9) and (3.38) where the coefficient b_7 remains as a free parameter. The value of b_7 has to be chosen in order that the set of values b_1, b_3, b_4, b_5, b_6 and b_7 do not take large values in order to prevent rounding off error when the solution is computed through formula (3.27). It remains to determine the value of the parameter δ.

Let's $u_k^{\mathrm{ip}}(s_{k+1}, \cdot)$ be the solution to problem (2.11) at grid point s_{k+1} and let's $\widetilde{u}_{k+1}^{\mathrm{ip},(4)}$ (resp. $\widetilde{u}_{k+1}^{\mathrm{ip},(5)}$) be the approximate solutions obtained from the 4th order (resp. 5th order) RK scheme

as given by formula (3.24) (resp. by formula (3.23)). For a sufficiently smooth function $\mathcal{G}_k$, a Taylor expansion about s_k lead to the following expressions for the local error at grid point s_k

$$u_k^{\rm ip}(s_{k+1},r) - \widetilde{u}_{k+1}^{\rm ip,(4)}(r) = h_k^5 \sum_{j=1}^{9} a_j^{(6)}\, D_{5,j} + \emptyset(h_k^6), \tag{3.42}$$

$$u_k^{\rm ip}(s_{k+1},r) - \widetilde{u}_{k+1}^{\rm ip,(5)}(r) = h_k^6 \sum_{j=1}^{20} \widehat{a}_j^{(6)}\, D_{5,j} + h_k^7 \sum_{j=1}^{48} \widehat{a}_j^{(7)}\, D_{6,j} + \emptyset(h_k^7) \tag{3.43}$$

where $D_{i,j}$ are the *elementary differentials* which are function of $\mathcal{G}_k$, $u_k^{\rm ip}(s_k,r)$ and s_k only and therefore depend only on the problem itself, whereas $a_j^{(6)}$ and $\widehat{a}_j^{(i)}$, $i=6,7$, are the order i truncation error coefficients and depend only on the RK formula. Comparison of the efficiency of RK schemes can be achieved by solving benchmark test problems, but formulae (3.42) and (3.43) makes it clear that the relative performances of 2 RK formulae depend on the ODE problem considered (through the elementary differentials) and for a same problem on the values of the physical parameters present in the equation. It is therefore customary to compare RK formulae of the same order by examining their truncation error coefficients since it is reasonable to hope that the relative size of the coefficients of the elementary differentials in relations (3.42) and (3.43) (i.e. the truncation error coefficients) will indicate how usually behaves the local error for the RK formula [28]. We will use this comparison criterion to determine the "best" possible ERK5(4) scheme among all the possible choices for the free parameter δ.

3.4.1. *Looking for an optimal ERK5(4) scheme.* Since the local extrapolation mode is adopted, one way of choosing the values of the remaining free parameter δ would be to make the truncation coefficients error "small" for the 5th order formula, *i.e.* to choose δ in order to minimize the quantity $\|\widehat{a}^{(6)}\|_2 = \left(\sum_{j=1}^{20} (\widehat{a}_j^{(6)})^2\right)^{1/2}$.

For the 5th order RK scheme defined by Butcher tableau (3.41) the truncation coefficients error $\widehat{a}_i^{(6)}$, $i=1,\ldots,20$ are [9, 10, 11]

$$\widehat{a}_1^{(6)} = \frac{1}{120}\sum_{i=1}^{7} \widehat{b}_i c_i^5 - \frac{1}{720} = 0$$

$$\widehat{a}_2^{(6)} = \frac{1}{6}\sum_{i,j=1}^{7} \widehat{b}_i c_i^3 a_{i,j} c_j - \frac{1}{72} = 0$$

$$\widehat{a}_3^{(6)} = \frac{1}{2}\sum_{i,j,k=1}^{7} \widehat{b}_i c_i a_{i,j} c_j a_{i,k} c_k - \frac{1}{48} = 0$$

$$\widehat{a}_4^{(6)} = \frac{1}{4}\sum_{i,j=1}^{7} \widehat{b}_i c_i^2 a_{i,j} c_j^2 - \frac{1}{72} = \frac{1}{5760}\,\frac{1+16\,\delta-12\,\delta^2-240\,\delta^3}{1-2\,\delta}$$

$$\widehat{a}_5^{(6)} = \frac{1}{2}\sum_{i,j,k=1}^{7} \widehat{b}_i a_{i,j} c_j^2 a_{i,k} c_k - \frac{1}{72} = \widehat{a}_4^{(6)}$$

$$\widehat{a}_6^{(6)} = \frac{1}{6}\sum_{i,j=1}^{7} \widehat{b}_i c_i a_{i,j} c_j^3 - \frac{1}{144} = \frac{1}{2880} - \frac{1}{1440}\,\delta$$

$$\widehat{a}_7^{(6)} = \frac{1}{24}\sum_{i,j=1}^{7} \widehat{b}_i a_{i,j} c_j^4 - \frac{1}{720} = 0$$

$$\widehat{a}_8^{(6)} = \frac{1}{2}\sum_{i,j,k=1}^{7} \widehat{b}_i c_i^2 a_{i,j} a_{j,k} c_k - \frac{1}{72} = -\frac{1}{2880} - \frac{1}{240}\,\delta + \frac{1}{48}\,\delta^2$$

$$\widehat{a}_9^{(6)} = \sum_{i,j,k,m=1}^{7} \widehat{b}_i a_{i,j} a_{i,k} c_k a_{j,m} c_m - \frac{1}{72} = \widehat{a}_8^{(6)}$$

$$\widehat{a}_{10}^{(6)} = \sum_{i,j,k=1}^{7} \widehat{b}_i c_i a_{i,j} c_j a_{j,k} c_k - \frac{1}{48} = 3\,\widehat{a}_6^{(6)}$$

$$\widehat{a}_{11}^{(6)} = \frac{1}{2}\sum_{i,j,k=1}^{7} \widehat{b}_i a_{i,j} c_j^2 a_{j,k} c_k - \frac{1}{120} = 0$$

$$\widehat{a}_{12}^{(6)} = \frac{1}{2}\sum_{i,j,k,m=1}^{7} \widehat{b}_i a_{i,j} a_{j,k} c_k a_{j,m} c_m - \frac{1}{240} = -15\,\widehat{a}_1^{(6)}$$

$$\widehat{a}_{13}^{(6)} = \frac{1}{2}\sum_{i,j,k=1}^{7} \widehat{b}_i c_i a_{i,j} a_{j,k} c_k^2 - \frac{1}{144} = -\frac{1}{1440}\,\frac{1+\delta-30\,\delta^2}{1-2\,\delta}$$

$$\widehat{a}_{14}^{(6)} = \frac{1}{2}\sum_{i,j,k=1}^{7} \widehat{b}_i a_{i,j} c_j a_{j,k} c_k^2 - \frac{1}{180} = -2\,\widehat{a}_4^{(6)}$$

$$\widehat{a}_{15}^{(6)} = \frac{1}{6}\sum_{i,j,k=1}^{7} \widehat{b}_i a_{i,j} a_{j,k} c_k^3 - \frac{1}{720} = -\widehat{a}_6^{(6)}$$

$$\widehat{a}_{16}^{(6)} = \sum_{i,j,k,m=1}^{7} \widehat{b}_i c_i a_{i,j} a_{j,k} a_{k,m} c_m - \frac{1}{144} = 15\,\widehat{a}_6^{(6)} - \frac{1}{360}$$

$$\widehat{a}_{17}^{(6)} = \sum_{i,j,k,m=1}^{7} \widehat{b}_i a_{i,j} c_j a_{j,k} a_{k,m} c_m - \frac{1}{180} = -2\,\widehat{a}_8^{(6)}$$

$$\widehat{a}_{18}^{(6)} = \sum_{i,j,k,m=1}^{7} \widehat{b}_i a_{i,j} a_{j,k} c_k a_{k,m} c_m - \frac{1}{240} = -3\,\widehat{a}_6^{(6)}$$

$$\widehat{a}_{19}^{(6)} = \frac{1}{2} \sum_{i,j,k,m=1}^{7} \widehat{b}_i a_{i,j} a_{j,k} a_{k,m} c_m^2 - \frac{1}{720} = -\widehat{a}_{13}^{(6)}$$

$$\widehat{a}_{20}^{(6)} = \sum_{i,j,k,m,n=1}^{7} \widehat{b}_i a_{i,j} a_{j,k} a_{k,m} a_{m,n} c_n - \frac{1}{720} = -15\,\widehat{a}_6^{(6)} + \frac{1}{360}$$

It follows that

$$\begin{aligned}\|\widehat{a}^{(6)}\|_2^2 &= \sum_{j=1}^{20} (\widehat{a}_j^{(6)})^2 \\ &= \frac{267 - 2416\,\delta + 11000\,\delta^2 - 43232\,\delta^3 + 120304\,\delta^4 - 224640\,\delta^5 + 345600\,\delta^6}{16588800\,(1-2\,\delta)^2}.\end{aligned}$$

Thus we are interested in finding the minimum value of $f(\delta) = \|\widehat{a}^{(6)}\|_2^2$ for $\delta \in]0, \frac{1}{2}[$. One can show that the absolute minimum of $\|\widehat{a}^{(6)}\|_2$ for the ERK5(4) scheme is attained in the set $]0, \frac{1}{2}[$ for an unique value $\delta_{\rm opt}$ approximately equal to 0.2370817283. For convenience, the rational approximation with 2 significant digits $1/4$ will be used. Figure 1 shows that this approximation is reasonable since $f(1/4) = 0.24\,10^{-5}$ whereas $f(\delta_{\rm opt}) = 0.21\,10^{-5}$. This choice for δ gives $\|\widehat{a}^{(6)}\|_2 = 1.54\,10^{-3}$. For comparison, Dormand and Prince RK5(4)-7M formula [24] gives $\|\widehat{a}^{(6)}\|_2 = 3.99\,10^{-4}$. We refer to [28] for a comparison with other classical ERK5(4) schemes.

With the choice of $b_7 = 1/14$ we finally obtain the following Butcher tableau for our ERK5(4) scheme

0							
1/2	1/2						
1/4	3/16	1/16					
1/2	−1/4	−1/4	1				
3/4	3/16	0	0	9/16			
1	−2/7	1/7	12/7	−12/7	8/7		
1	7/90	0	16/45	2/15	16/45	7/90	
	1/14	0	8/21	2/21	8/21	0	1/14

(3.44)

This ERK5(4) scheme belongs to the family of *Quadrature Defective Runge-Kutta methods* since it has 2 quadrature nodes c_2 and c_4 equals. It means that the RK approximation formula does not coincide with a quadrature rule when solving an ODE in the form $y'(t) = F(t, y(t))$ with F function of t only such as does the classical RK4 formula which coincides in such a case with Simpson's quadrature rule.

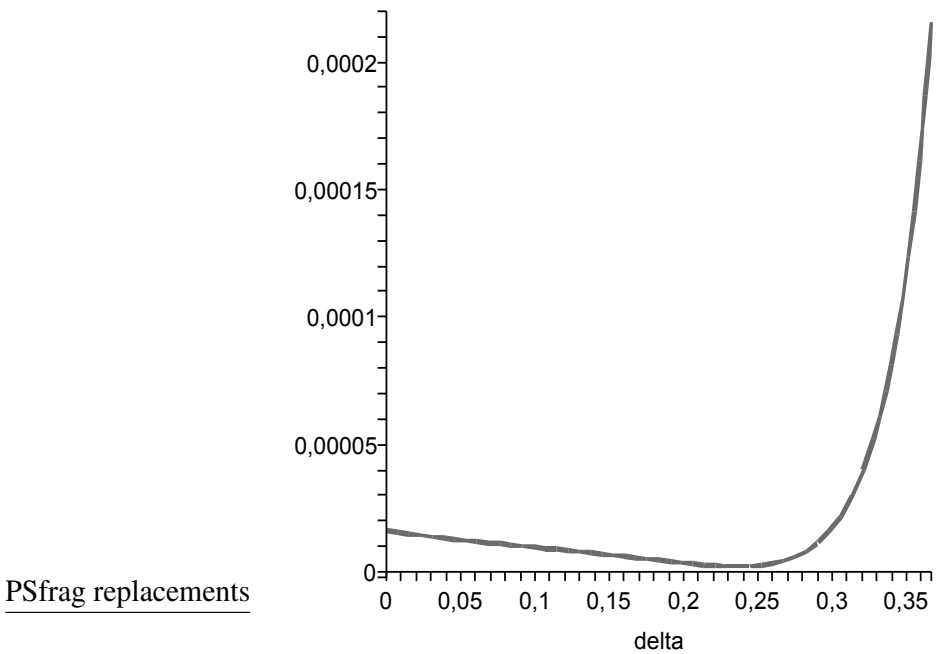

FIGURE 1. Graph of the mapping $f : \delta \mapsto \sum_{j=1}^{20} (\widehat{a}_j^{(6)})^2$ over $[0, \frac{1}{2}]$.

3.4.2. *Stability domain of the ERK5(4) scheme.* As well known when the incremental function of an explicit RK formula satisfies a condition of Lipschitz type, the RK method is *stable* provide the step-size h is small enough, meaning that small changes in the initial data or in the ODE produce bound changes in the numerical solution in the limit case when h tends to 0 (see e.g. [9, 10, 11] for the exact definition). The notion of *absolute stability* has been introduced to provide a more practical tool when studying RK methods especially with regard to the small step-size limit validity condition when conversely with adaptive step-size strategies we are interested in steps with the maximal possible size to meet a given accuracy. A numerical method is said to be absolutely stable for a step-size h and a given ODE if a change in the initial data of size ϵ_0 is no larger than ϵ_0 in the all subsequent steps, see e.g. [9, 10, 29]. Since the definition of absolute stability depends on the ODE, it is common to study it on the linear test problem: $y'(t) = \lambda\, y(t)$ subject to the initial condition $y(0) = y_0$ for $\lambda \in \mathbb{C}$, $\Re(\lambda) \leq 0$. The region of absolute stability of a RK scheme is the set of all non-negative values of h and complex values λ for which the RK scheme is absolutely stable when applied to the linear test problem. RK methods can be compared on the basis of the size of their region of absolute stability. A 5th order RK formula with optimal region of stability is propound in [30].

When the 4th and 5th order RK formulae defined in the extend Butcher tableau (3.44) are used to solve the above mentioned linear test problem we obtain for the solution at grid point t_k,

$k \in \mathbb{N}$ the following approximation formulae $y_{k+1}^{(4)} = R^{(4)}(\lambda h)\, y_k^{(4)}$ and $y_{k+1}^{(5)} = R^{(5)}(\lambda h)\, y_k^{(5)}$ where

$$R^{(4)}(z) = 1 + z + \frac{1}{2}\, z^2 + \frac{1}{6}\, z^3 + \frac{1}{24}\, z^4 + \frac{1}{120}\, z^5 + \frac{1}{640}\, z^6,$$
$$R^{(5)}(z) = 1 + z + \frac{1}{2}\, z^2 + \frac{1}{6}\, z^3 + \frac{1}{24}\, z^4 + \frac{13}{1344}\, z^5 + \frac{1}{1680}\, z^6 + \frac{1}{8960}\, z^7.$$

Thus, a change in the initial data y_0 of size ϵ_0 produces a deviation at step k of size $\epsilon_k = (R^{(p)}(\lambda h))^{k-1}\, \epsilon_0$ (for $p = 4, 5$). Thus the initial perturbation will not grow beyond ϵ_0 if $|R^{(p)}(\lambda h)| \leq 1$. The region of stability of the 4th and 5th order RK formulae is then defined as the set of complex number z with a negative real part such that $|R^{(p)}(z)| \leq 1$. They are depicted in figure 2. We obtain regions of stability with a size very similar to the ones of the Dormand and Price RK5(4)7M formula [24].

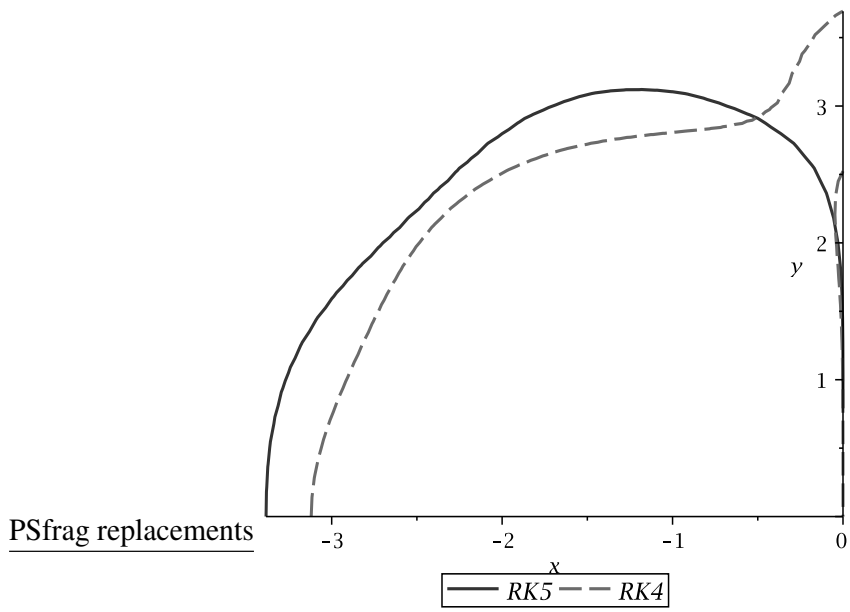

FIGURE 2. Stability domain in the complex plane of the 4th and 5th order RK formulae (the regions are symmetric with respect to the real axis).

3.4.3. *Algorithm for the ERK5(4)-IP method.* The computational sequence for one step of the ERK5(4) scheme is now rewritten with the coefficient values given in Butcher tableau (3.41) and optimized to reduce the number of exponential operator terms to be evaluated. The approximate solution $\widetilde{u}_{k+1}^{(5)}$ to problem (2.9) at grid point s_{k+1} is obtained from the approximate value $\widetilde{u}_k^{(5)}$ at grid point s_k by the following computational sequence:

$$\widetilde{u}_k^{\mathrm{ip},(5)}(r) = \mathrm{e}^{\frac{h_k}{2}\mathcal{D}} \cdot \widetilde{u}_k^{(5)}(r)$$

$$
\begin{aligned}
\alpha_1 &= \mathrm{e}^{\frac{h_k}{2}\mathcal{D}} \cdot \alpha'_{7,k} \quad \text{where } \alpha'_{7,k} \text{ was computed at the previous step}\\
\alpha_2 &= \mathcal{N}\left(\widetilde{u}_k^{\mathrm{ip},(5)}(r) + \frac{h_k}{2}\alpha_1\right)\\
\alpha_3 &= \mathrm{e}^{\frac{h_k}{4}\mathcal{D}} \cdot \mathcal{N}\Big(\mathrm{e}^{-\frac{h_k}{4}\mathcal{D}} \cdot \big[\widetilde{u}_k^{\mathrm{ip},(5)}(r) + \frac{h_k}{16}(3\alpha_1 + \alpha_2)\big]\Big)\\
\alpha_4 &= \mathcal{N}\Big(\widetilde{u}_k^{\mathrm{ip},(5)}(r) + \frac{h_k}{4}(\alpha_1 - \alpha_2 + 4\alpha_3)\Big)\\
\alpha_5 &= \mathrm{e}^{-\frac{h_k}{4}\mathcal{D}} \cdot \mathcal{N}\Big(\mathrm{e}^{\frac{h_k}{4}\mathcal{D}} \cdot \big[\widetilde{u}_k^{\mathrm{ip},(5)}(r) + \frac{3h_k}{16}(\alpha_1 + 3\alpha_4)\big]\Big)\\
\alpha'_6 &= \mathcal{N}\Big(\mathrm{e}^{\frac{h_k}{2}\mathcal{D}} \cdot \big[\widetilde{u}_k^{\mathrm{ip},(5)}(r) + \frac{h_k}{7}(-2\alpha_1 + \alpha_2 + 12\alpha_3 - 12\alpha_4 + 8\alpha_5)\big]\Big)\\
\widetilde{u}_{k+1}^{(5)} &= \mathrm{e}^{\frac{h_k}{2}\mathcal{D}} \cdot \big(\widetilde{u}_k^{\mathrm{ip},(5)}(t) + \frac{h_k}{90}(7\alpha_1 + 32\alpha_3 + 12\alpha_4 + 32\alpha_5)\big) + \frac{7h_k}{90}\alpha'_6\\
\alpha'_{7,k+1} &= \mathcal{N}\Big(\widetilde{u}_{k+1}^{(5)}\Big)\\
\widetilde{u}_{k+1}^{(4)} &= \mathrm{e}^{\frac{h_k}{2}\mathcal{D}} \cdot \big(\widetilde{u}_k^{\mathrm{ip},(5)}(t) + \frac{h_k}{42}(3\alpha_1 + 16\alpha_3 + 4\alpha_4 + 16\alpha_5)\big) + \frac{h_k}{14}\alpha'_{7,k+1}
\end{aligned}
$$

Compared to the initial computational sequence, this reformulation saves the computation of the $\exp(-\frac{h_k}{2}\mathcal{D})$ term involved in the expression of α_6 and α_7 since a cancellation happens with the $\exp(\frac{h_k}{2}\mathcal{D})$ term in the expression of $\widetilde{u}_{k+1}^{(4)}$ and $\widetilde{u}_{k+1}^{(5)}$. Moreover, although the ERK5(4) scheme appears as a 7 stage method, its effective cost is very similar to a 6 stage method since the computation of the first coefficient α_1 at step $k+1$ uses the results of $\alpha'_{7,k}$ computed at step k. Compared to the ERK4(3) scheme for the IP method presented in [14], the propound computational procedure requires 6 evaluations of the nonlinear operator $\mathcal{N}$ instead of 4 and it requires the additional computation of the 2 exponential operators $\mathrm{e}^{\frac{h_k}{4}\mathcal{D}}$ and $\mathrm{e}^{-\frac{h_k}{4}\mathcal{D}}$ for a benefit corresponding to the gain of one convergence order.

3.4.4. *Local error estimate.* The local error at step k can be estimated from the values $\widetilde{u}_{k+1}^{(4)}$ and $\widetilde{u}_{k+1}^{(5)}$ as follows. Assuming enough regularity on the solution functions, the local errors at grid point s_{k+1} for the RK4 and the RK5 formulae are respectively given by [9, 10, 11]: $\forall r \in \Omega$

$$
\begin{aligned}
\ell_{k+1}^{(4)}(r) &= u(s_{k+1}, r) - \widetilde{u}_{k+1}^{(4)}(r) = \psi_4(s_k, r, \widetilde{u}_k^{(4)})\, h_k^5 + \emptyset(h_k^6),\\
\ell_{k+1}^{(5)}(r) &= u(s_{k+1}, r) - \widetilde{u}_{k+1}^{(5)}(r) = \psi_5(s_k, r, \widetilde{u}_k^{(5)})\, h_k^6 + \emptyset(h_k^7),
\end{aligned}
\tag{3.45}
$$

where ψ_4 and ψ_5 are functions of the elementary differential of order 4 and 5 respectively. By difference of these 2 relations we obtain

$$
\widetilde{u}_{k+1}^{(5)}(r) - \widetilde{u}_{k+1}^{(4)}(r) = \psi_4(s_k, r, \widetilde{u}_k^{(4)})\, h_k^5 + \emptyset(h_k^6).
$$

Thus the local error for the 4th order RK formula at grid point s_{k+1} can be approximated, with an error in $\emptyset(h_k^5)$, in the following way:

$$\forall r \in \Omega \qquad \ell_{k+1}^{(4)}(r) = \psi_4(s_k, r, \widetilde{u}_k^{(4)})\, h_k^5 + \emptyset(h_k^6) \approx \widetilde{u}_{k+1}^{(5)}(r) - \widetilde{u}_{k+1}^{(4)}(r). \tag{3.46}$$

The quadratic local error at grid point s_{k+1} is then

$$L_{k+1}^{(4)} = \|\ell_{k+1}^{(4)}\|_{\mathbb{L}^2(\Omega,\mathbb{C})} \approx \left(\int_\Omega \left| \widetilde{u}_{k+1}^{(5)}(r) - \widetilde{u}_{k+1}^{(4)}(r) \right|^2 \mathrm{d}r \right)^{\frac{1}{2}}, \tag{3.47}$$

where the integral has to be evaluated by means of a quadrature formula. As mentioned before, even if the local error estimate (3.46) holds only for the 4th order method, when the local extrapolation mode is used the value given by the 5th order formula as the approximation of the solution at grid point s_{k+1} is propagated. In general this approach overestimates the actual local error, which is safe but not optimal.

4. Numerical experiments

In the framework of a project on the numerical simulation of incoherent optical wave propagation in nonlinear fibers [8] we have implemented the ERK5(4)-IP method for solving the GNLS problem (2.2). We present in this section numerical results from the ERK5(4)-IP method on 2 selected applications in optics: the propagation of optical solitons and the propagation of a picosecond pulse into a single-mode fiber where fiber losses, nonlinear Raman and Kerr effects as well as high order chromatic dispersion are taken into account. In both cases, the adaptive step-size strategy using the ERK5(4) scheme is compared to the ERK4(3) scheme for the IP method presented in [14] and to the one based on the step-doubling (SD) approach [31].

4.1. **Soliton solution to the NLSE in optics.** We first consider the case of the nonlinear Schrödinger equation (NLSE) in optics, a simplified version of the GNLSE (2.2) where $\alpha = 0$, $f_R = 0$, $n_{\max} = 2$. The linear operator is $\mathcal{D} : A \mapsto \mathrm{i}\beta_2 \partial_{tt} A$ and the nonlinear operator is $\mathcal{N} : A \mapsto \mathrm{i}\gamma A \left|A\right|^2$. When $\beta_2 < 0$, there exists an exact solution to the NLSE known as the optical soliton [17]. Namely, if the source term is given by

$$\forall t \in \mathbb{R} \quad a_0(t) = \frac{N}{\sqrt{\gamma L_D}} \frac{1}{\cosh(Nt/T_0)} \tag{4.1}$$

where N is the soliton order, T_0 is the pulse half-width and $L_D = -T_0^2/\beta_2$ is the dispersion length then the solution to the NLSE at any position $z \in [0, L]$ along the fiber reads

$$\forall t \in \mathbb{R} \quad A(z,t) = \frac{N}{\sqrt{\gamma L_D}} \frac{\mathrm{e}^{\mathrm{i}zN^2/2L_D}}{\cosh(Nt/T_0)}. \tag{4.2}$$

Fundamental soliton ($N = 1$) doesn't provide a well suited example for exploring the features of the ERK5(4)-IP method and for comparison purposes since its shape doesn't change on propagation. We therefore consider in the following a 3rd order soliton ($N = 3$). In Fig. 3 we show for the 3rd order soliton the adjustment of the step-size when using the ERK5(4)-IP method for evaluating the local error with a tolerance set to $\mathrm{tol} = 10^{-6}$ and an initial step-size

of $h = 1\,\mathrm{m}$. The other physical parameters of the numerical experiment are $L = 637.21\,\mathrm{m}$, $\gamma = 4.3\,\mathrm{W}^{-1}\,\mathrm{km}^{-1}$, $\beta_2 = -19.83\,\mathrm{ps}^2\,\mathrm{km}^{-1}$, $T_0 = 2.8365\,\mathrm{ps}$. The number of discretisation steps along the fiber is found to be 454 and the computation time is $72\,\mathrm{s}$ on a Intel Core 2 Quad Q6600. At the fiber end ($z = L$), the relative global error measured with the quadratic norm is $5.53\,10^{-5}$ whereas the maximum relative error is $9.84\,10^{-5}$.

The same accuracy with a constant step-size computation would have required a step size of $0.01\,\mathrm{m}$ for a total number of step of 63722 and a computation CPU time of $5490\,\mathrm{s}$. For comparison, when using the ERK4(3) scheme, the number of discretisation steps along the fiber is found to be 605 and the computation time is $69\,\mathrm{s}$; the relative global error at the fiber end measured with the quadratic norm is $1.12\,10^{-4}$ whereas the maximum relative error is $1.89\,10^{-4}$. When using an adaptive step-size strategy based on the SD approach with the same values of tolerance and initial step-size, we obtain that the number of discretisation steps along the fiber is 396 (or 792 if we consider that it is the accurate solution computed over the fine grid of step-size $h_k/2$ that is propagated) and the computation time is $148\,\mathrm{s}$. At the fiber end ($z = L$), the relative quadratic error is $8.83\,10^{-6}$ whereas the maximum relative error is $1.48\,10^{-5}$. The evolution of the step-size along the fiber is depicted in Fig. 3 for a comparison with the ERK5(4)-IP method. We can see that whatever is the method used for estimating the local error we obtain a very similar shape for the adaptive step-size curve. The only difference lies in the size of the steps which are larger with the ERK5(4) scheme for the same value of the tolerance prescribed for the local error. Namely, when comparing the step-size strategy for the ERK5(4) scheme to the one of the ERK4(3) scheme we can see that for a comparable computational time and accuracy of the results, the steps are increased by an average ration of approximately $4/3$.

4.2. **Solving the GNLSE in optics by the ERK5(4) method.** We now consider the case of the GNLSE (2.2) with the following set of physical parameters : $\omega_0 = 1770\,\mathrm{Thz}$, $\gamma = 4.3\,\mathrm{W}^{-1}\mathrm{km}^{-1}$, $\beta_2 = 19.83\,\mathrm{ps}^2\mathrm{km}^{-1}$, $\beta_3 = 0.031\,\mathrm{ps}^3\mathrm{km}^{-1}$ and $\beta_n = 0$ for $n \geq 4$, $\alpha = 0.046\,\mathrm{km}^{-1}$, $L = 96.77\,\mathrm{m}$, $f_R = 0.245$. An expression for the Raman time response function h_R for silica core fiber is given in [17]. The Gaussian pulse at the fiber entrance ($z = 0$) is expressed as

$$\forall t \in \mathbb{R} \quad a_0(t) = \sqrt{P_0}\,\mathrm{e}^{-\frac{1}{2}(t/T_0)^2}, \tag{4.3}$$

where $T_0 = 2.8365\,\mathrm{ps}$ is the pulse half-width and $P_0 = 100\,\mathrm{W}$ is the pulse peak power.

In Fig. 4 we show the adjustment of the step-size when using the ERK5(4) scheme for evaluating the local error with a tolerance set to $\mathrm{tol} = 10^{-6}$ and an initial step size of $h = 0.1\,\mathrm{m}$. The number of discretisation steps along the fiber is found to be 170 and the computation time is $49.2\,\mathrm{s}$. For a comparison, when using the ERK4(3) scheme, the number of discretisation steps along the fiber is found to be 279 and the computation time is $50\,\mathrm{s}$. When using the SD method for determining the step-size in the IP method in the same circumstances we find that the number of discretisation steps along the fiber is 232 and the computation time is $124\,\mathrm{s}$. The same comments as for the soliton case can be made when comparing the adaptive step-size approaches.

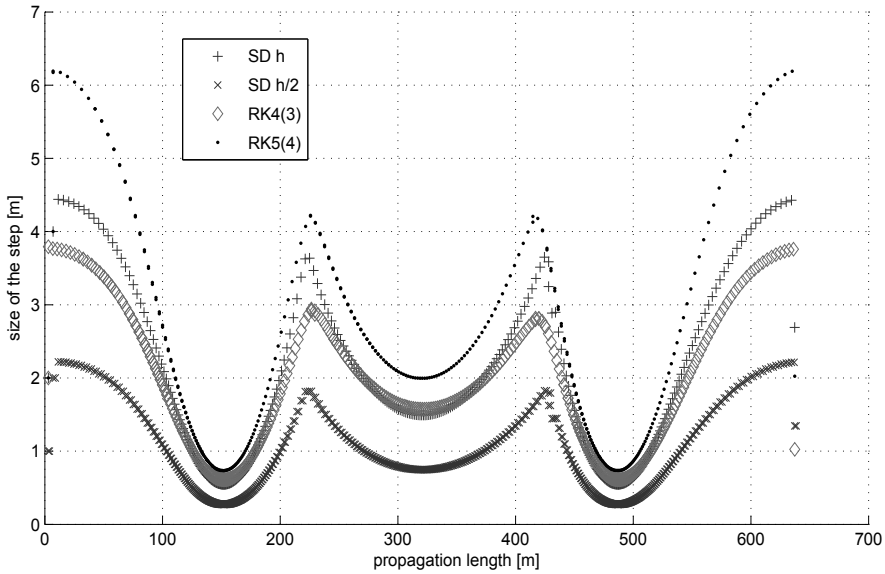

FIGURE 3. Evolution of the step-size along the fiber length for adaptive step-size strategy based on the ERK5(4), ERK4(3) and SD methods (considered over the coarse and fine grids) when solving the NLSE for a 3rd order soliton.

When the tolerance is set to $\mathrm{tol} = 10^{-9}$ with an initial step size of $h = 0.1\,\mathrm{m}$, the number of step-size with the ERK5(4)-IP method is 671 and the computational time is $177\,\mathrm{s}$ whereas 1545 steps are required by the RK4(3) method for a computational time of $221\,\mathrm{s}$ and 906 steps are required by the SD method for a computational time of $645\,\mathrm{s}$. The evolution of the step-size along the fiber length is very similar to the one presented in Fig. 4.

5. CONCLUSION

In this paper we have presented a 5th order Runge-Kutta (RK) formula with 6 computational stages designed to be use in conjunction with the IP method in the sense that the coefficients of the 5th order RK formula have been determined in order to reduce the global computational cost of the IP method. Moreover, we have designed the 5th order RK scheme so that it is embedded in a 4th order RK scheme with 7 computational stages in order to dispose of local error estimations for adaptive step-size control purposes in the IP method. Compared to the embedded RK pair of order 4 and 3 (ERK4(3) scheme) for the IP method presented in [14], the propound computational procedure (ERK5(4)-IP method) requires 2 additional evaluations of the nonlinear operator as well as 2 additional evaluations of the exponential operator. However the benefit of this higher order embedded RK scheme for the IP method is to deliver results at a certain accuracy with less computational steps than with the ERK4(3) scheme and therefore the method is likely to reduce the accumulation of round-off errors. The numerical experiments

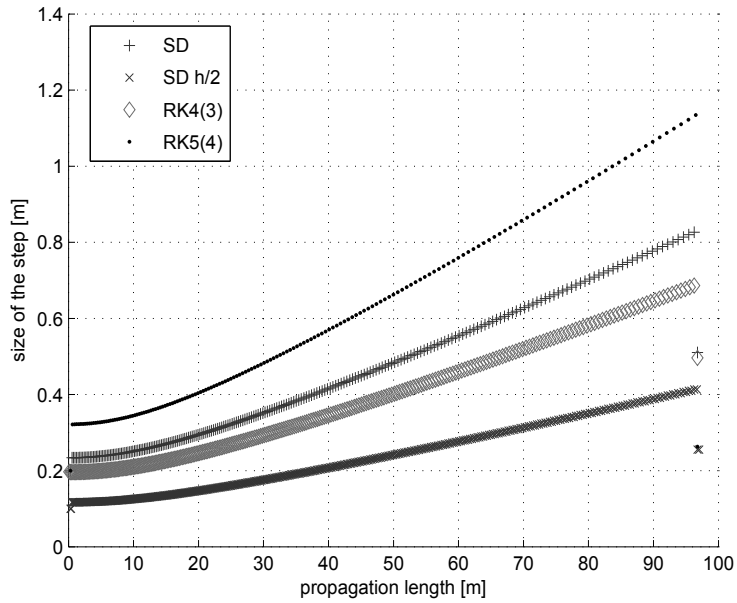

FIGURE 4. Evolution of the step-size along the fiber length for adaptive step-size strategy based on the ERK5(4), ERK4(3) and SD methods (considered over the coarse and fine grids) when solving the GNLSE.

we have conducted show that compared to the ERK4(3)-IP method, the ERK5(4)-IP method provides, for a computational time very similar, sightly more accurate results with a much lower number of discretization steps. Therefore, as expected, the additional computational cost per step of the ERK5(4) scheme compared to the ERK4(3) scheme is counterbalance by the lower number of steps required.

As mentioned in the text, since the IP method and the Symmetric Split-step (SS) method have a very similar internal computational structure, the propound ERK5(4) scheme could be used in conjunction with the SS method for solving the GNLSE with the same advantages as the one mentioned above (apart an additional error due to the use of a splitting formula).

ACKNOWLEDGMENTS

This work has been undertaken under the framework of the Green-Laser project and was partially supported by Conseil Régional de Bretagne, France. The author would like to thank F. Mahé and R. Texier-Picard from the Institute of Mathematics (IRMAR CNRS UMR 6625) in Rennes, France, for their involvement in the study of the IP method as well as A. Fernandez from LAAS (UPR CNRS 8001) in Toulouse, France, for his major contribution to the Green-Laser project.

REFERENCES

[1] B.M. Caradoc-Davies. *Vortex dynamics in Bose-Einstein condensate*. PhD thesis, University of Otago (NZ), 2000.

[2] M.J. Davis. *Dynamics in Bose-Einstein condensate*. PhD thesis, University of Oxford (UK), 2001.

[3] S. Wüster, T.E. Argue, and C.M. Savage. Numerical study of the stability of skyrmions in Bose-Einstein condensates. *Phys. Rev. A*, 72(4), 2005.

[4] R. Scott, C. Gardiner, and D. Hutchinson. Nonequilibrium dynamics: Studies of the reflection of Bose-Einstein condensates. *Laser Phys.*, 17:527–532, 2007.

[5] C.N. Liu, G.G. Krishna, M. Umetsu, and S. Watanabe. Numerical investigation of contrast degradation of Bose-Einstein condensate interferometers. *Phys. Rev. A*, 79(1), 2009.

[6] J. Hult. A fourth-order Runge–Kutta in the Interaction Picture method for simulating supercontinuum generation in optical fibers. *J. Lightwave Technol.*, 25(12):3770–3775, 2007.

[7] A. Heidt. Efficient adaptive step size method for the simulation of supercontinuum generation in optical fibers. *J. Lightwave Technol.*, 27(18):3984–3991, 2009.

[8] A. Fernandez, S. Balac, A. Mugnier, F. Mahé, R. Texier-Picard, T. Chartier, and D. Pureur. Numerical simulation of incoherent optical wave propagation in nonlinear fibers. *To appear in Eur. Phys. J. - Appl. Phys.*, 2013.

[9] J.C. Butcher. *Numerical methods for ordinary differential equations*. John Wiley and Sons, 2008.

[10] E. Hairer, S.P. Norsett, and G. Wanner. *Solving ordinary differential equations I: nonstiff problems*. Springer-Verlag, 1993.

[11] M. Crouzeix and A. Mignot. *Analyse numérique des équations différentielles*. Masson, Paris, 1984.

[12] J.S. Townsend. *A modern approach to quantum mechanics*. International series in pure and applied physics. University Science Books, 2000.

[13] M. Guenin. On the interaction picture. *Commun. Math. Phys.*, 3:120–132, 1966.

[14] S. Balac and F. Mahé. Embedded Runge-Kutta scheme for step-size control in the Interaction Picture method. *Comput. Phys. Commun.*, 184:1211–1219, 2013.

[15] S. N. Papakostas and G. Papageorgiou. A family of fifth-order RungeKutta pairs. *Math. Comp*, 65:215, 1996.

[16] J.R. Cash and A.H. Karp. A variable order Runge-Kutta method for initial value problems with rapidly varying right-hand sides". *ACM Trans. Math. Software*, 16:201–222, 1990.

[17] G. Agrawal. *Nonlinear fiber optics*. Academic Press, 3rd edition, 2001.

[18] B.M. Caradoc-Davies, R.J. Ballagh, and P.B. Blakie. Three-dimensional vortex dynamics in Bose-Einstein condensates. *Phys. Rev. A*, 62:011602, 2000.

[19] A. Pazy. *Semigroups of Linear Operators and Applications to Partial Differential Equations*. Number vol. 44 in Applied Mathematical Sciences. Springer, 1992.

[20] S. Balac, A. Fernandez, F. Mahé, F. Méhats, and R. Texier-Picard. The Interaction Picture method for solving the Generalized Nonlinear Schrödinger Equation in optics. *submitted to SIAM J. Numer. Anal.*, 2013.

[21] G. Strang. On the construction and comparison of difference schemes. *SIAM J. Numer. Anal.*, 5(3):506–517, 1968.

[22] J.C. Butcher. On Runge-Kutta processes of high order. *J. Aust. Math. Soc.*, 4(02):179–194, 1964.

[23] E. Fehlberg. Low order classical Runge-Kutta formulas with stepsize control and applications to some heat transfert problems. Technical report, National Aeronautics and Space Administration, 1969.

[24] J.R. Dormand and P.J. Prince. A family of embedded Runge-Kutta formulae. *J. Comput. Appl. Math.*, 6:19–26, 1980.

[25] W. Kutta. Beitrag zur naherungsweisen integration totaler differentialgleichungen. *Z. Math. Phys.*, (46):434–453, 1901.

[26] C.R. Cassity. The complete solution of the fifth order Runge–Kutta equations. *SIAM J. Numer. Anal.*, 6(3):432–436, 1969.

[27] Maplesoft. *Maple 16 Programming Guide*. Waterloo Maple Inc., 2012.

[28] L. Shampine. Some practical Runge-Kutta formulas. *Math. Comp.*, 46:135–150, 1986.
[29] J.H.E. Cartwright and O. Piro. The dynamics of Runge-Kutta methods. *Int. J. Bifurcation and Chaos*, 2:427–49, 1992.
[30] J.D. Lawson. An order five Runge–Kutta process with extended region of stability. *SIAM J. Numer. Anal.*, 3(4):593–597, 1966.
[31] L. Shampine. Local error estimation by doubling. *Computing*, 34:179–190, 1985.