Runge–Kutta projection methods with low dispersion and dissipation errors

M. Calvo [1], M.P. Laburta [1], J.I. Montijano[1] and L. Rández[1] *

November 25, 2013

[1] *IUMA – Departamento de Matemática Aplicada*
Universidad de Zaragoza. 50009-Zaragoza, Spain.
email: calvo, laburta, monti, randez@unizar.es

Abstract

In this paper new one-step methods that combine Runge–Kutta (RK) formulae with a suitable projection after the step are proposed for the numerical solution of Initial Value Problems. The aim of this projection is to preserve some first integral in the numerical integration. In contrast with standard orthogonal projection, the direction of the projection at each step is obtained from another suitable embedded formula so that the overall method is affine invariant. A study of the local errors of these projection methods is carried out, showing that by choosing proper embedded formulae the order can be increased for the harmonic oscillator. Particular embedded formulae for the third order method by Bogacki and Shampine (BS3) are provided. Some criteria to get appropriate dynamical directions for general problems as well as sufficient conditions that ensure the existence of RK methods embedded in BS3 according to them are given. Finally, some numerical experiments to test the behaviour of the new projection methods are presented.

AMS subject classification: 65L05, 65L06.

Keywords: initial value problems, numerical geometric integration, projection methods, dispersion error, explicit Runge–Kutta methods.

*This work was supported by D.G.I. project MTM2010-21630-C02-01.

1 Introduction.

We consider autonomous systems of ordinary differential equations

$$y'(t) = f(y(t)), \tag{1}$$

with a sufficiently smooth N-dimensional vector field $f\colon D \subseteq \mathbb{R}^N \to \mathbb{R}^N$, having a scalar invariant $G(y)$.

As it is well known [16, pp. 93], a scalar function $G\colon \widehat{D} \subseteq \mathbb{R}^N \to \mathbb{R}$ of class $\mathcal{C}^{(1)}(\widehat{D})$, $\widehat{D} \subseteq D$, is a first integral in $\widehat{D}$ (also called strong invariant or conserved quantity) of (1) iff

$$\nabla G(y) \cdot f(y) = 0, \ \forall y \in \widehat{D},$$

which implies that every solution $y(t)$ contained in $\widehat{D}$ of (1) satisfies

$$G(y(t)) = G(y(t_0)) \quad \forall t.$$

In this case, for all $y_0 \in \widehat{D}$, the solution $y = y(t)$ of (1) satisfying $y(t_0) = y_0$ is contained in the hypersurface

$$M_{y_0} = \{y \in \mathbb{R}^N \,|\, G(y) = G(y_0)\}. \tag{2}$$

For problems possessing a first integral, it is natural to ask whether or not a numerical method provides approximate solutions that stay in the hypersurface as the true solution does.

Energy-preserving Runge–Kutta methods for polynomial Hamiltonian dynamical systems were given in [21] and [19]. Hamiltonian Boundary Value Methods [4, 5] and the methods presented in [20] preserve the energy also for polynomial Hamiltonians. In [10], [15] and [23], energy preserving methods for general Hamiltonian systems were proposed. The preservation of the energy for Poisson systems was studied also in [2] and [11].

Quispel and Capel [22] proposed discrete algorithms to preserve exactly general first integrals. On the other hand, *Line Integral Methods*, where the key idea of imposing energy conservation through a line integral has been extended to any invariant and to any conservative problem, were introduced in [3].

All these preserving methods are in general implicit. Explicit RK methods are able to preserve only linear first integrals even though quadratic invariants can be approximated very accurately by pseudo symplectic methods [8]. A natural way to obtain one-step numerical methods that preserve a given invariant based on explicit methods is to start with a standard scheme and add a suitable projection. Projection methods can also be used to preserve monotonicity of Lyapunov functions [9, 14].

Let $y(t)$ be the solution of (1) satisfying $y(t_0) = y_0$. In order to obtain numerical methods that preserve first integrals of (1), projection methods

provide, with a step size h, approximations y_n to $y(t_n)$, with $t_n = t_0 + nh$, of the form

$$y_{n+1} = \widetilde{y}_{n+1} + \lambda_n w_n, \ n = 0, 1, 2, \ldots \tag{3}$$

where

- $\widetilde{y}_{n+1}$ is the numerical approximation to $y(t_{n+1})$ given by a standard (non-preserving) method.
- $w_n \in \mathbb{R}^N$ is the vector (depending on y_n and the step size h) that defines the direction of the projection.
- λ_n is a real parameter which, once determined the direction vector w_n, will be calculated so that y_{n+1} is the projection of $\widetilde{y}_{n+1}$ onto the variety (2). So, denoting $g(y) := G(y) - G(y_0)$, λ_n will be computed at each step so that

$$g(y_{n+1}) = g(\widetilde{y}_{n+1} + \lambda_n w_n) = 0.$$

In this paper we will focus on projection methods applied to Runge–Kutta (RK) schemes. This means that $\widetilde{y}_{n+1}$ is the approximation provided by an s-stage explicit RK method with coefficients $(A, \widetilde{b})$, with $A = (a_{ij}) \in \mathbb{R}^{s\times s}$, and $\widetilde{b} = (\widetilde{b}_i) \in \mathbb{R}^s$.

Standard orthogonal projection methods [16, p. 106] are of the form (3) with $w_n = \nabla g(y_{n+1})$. In [16], this vector w_n is replaced by $w_n = \nabla g(\widetilde{y}_{n+1})$ to reduce the computational cost. One drawback of these projection techniques is that, whereas RK methods preserve linear first integrals [12], after applying either standard or simplified orthogonal projection, this does not hold. Another inconvenience is that the projected method obtained with orthogonal projections is not affine invariant like the basic RK formula is. Thus, the behaviour of the numerical method integrating linear differential systems $y' = Ly$, with L diagonalizable, is not equivalent to the behaviour when the diagonal system is integrated.

In [7], the authors propose to search for a dynamical direction w_n at each step so that the projected method is also a RK method. In order to do that, they take

$$w_n = \widehat{y}_{n+1} - \widetilde{y}_{n+1}, \tag{4}$$

where $\widehat{y}_{n+1}$ is given by an explicit RK method with coefficients $(A, \widehat{b})$, $\widehat{b} = (\widehat{b}_i) \in \mathbb{R}^s$, embedded to the RK method that gives the approximation $\widetilde{y}_{n+1}$. Then,

$$y_{n+1} = (1 - \lambda_n)\widetilde{y}_{n+1} + \lambda_n \widehat{y}_{n+1}, \tag{5}$$

and y_{n+1} is a convex linear combination of RK approximations and so, it is also a RK method. Such a projected RK method y_{n+1} with direction vector based on an embedded RK pair preserves all linear first integrals and it is also affine invariant [7].

The best way to choose the embedded method $(A, \widehat{b})$ is, to our knowledge, an open question, and this paper tries to gain some insight into this problem. The orthogonal projection chooses the direction looking for the approximation into the manifold $y_{n+1} \in M_{y_0}$ closest to the approximation $\widetilde{y}_{n+1}$, that is, it minimizes the distance $\|\widetilde{y}_{n+1} - y_{n+1}\|$. However, this criterion, which is usual in projection techniques, might not be the best one in our problem. It would be more convenient to minimize the distance with respect to the true solution $\|y(t_{n+1}) - y_{n+1}\|$. Unfortunately, the true solution is not known.

In section 2 we give some results about the order that the projected method (5) achieves depending on how the direction w_n given in (4) is chosen. Section 3 is focused on oscillatory problems, showing that it is possible to get zero-dissipative RK projected methods with high order for the harmonic oscillator. In particular, the orders of some projection methods based on the third order Bogacki–Shampine RK method (BS3) are studied. In Section 4 we establish some criteria to get appropriate directions w_n for general problems, and we give some conditions under which those criteria are satisfied when BS3 is projected. Finally, in Section 5 we present some numerical experiments to check the theoretical results and to show the efficiency of the new projection methods.

2 The local error of RK projection methods

The properties of the projected solution y_{n+1} in (3) depend on the direction w_n, and in the case of a Runge–Kutta projection this direction vector depends on the embedded method $\widehat{y}_{n+1}$. We are interested in the best way of choosing the free coefficients $\widehat{b}_i$ so that the projected solution is as close to the true solution as possible.

One can ask if it is possible to select the free parameters of the embedded method so that the projected solution increases the order with respect to the non-projected solution, or it has at least a leading term with smaller coefficients. Let us first analyze the local error of the projected solution.

Let us suppose that the non-projected method has order p and the embedded method has order $q < p$, and their local errors admit expansions

$$y(t_n + h; t_n, y_n) - \widetilde{y}_{n+1} = h^{p+1}\widetilde{\varphi}_{p+1}(y_n) + h^{p+2}\widetilde{\varphi}_{p+2}(y_n) + \mathcal{O}(h^{p+3}),$$
$$y(t_n + h; t_n, y_n) - \widehat{y}_{n+1} = h^{q+1}\widehat{\varphi}_{q+1}(y_n) + h^{q+2}\widehat{\varphi}_{q+2}(y_n) + \mathcal{O}(h^{q+3}),$$

where $y(t; t_n, y_n)$ represents the local solution of the differential system that satisfies $y(t_n; t_n, y_n) = y_n$.

The direction of projection $w_n = \widehat{y}_{n+1} - \widetilde{y}_{n+1} = \mathcal{O}(h^{q+1})$ admits an asymptotic expansion

$$w_n = h^{q+1}\psi_{q+1} + h^{q+2}\psi_{q+2} + \mathcal{O}(h^{q+3}),$$

with $\psi_{q+1} = -\widehat{\varphi}_{q+1}$.

We can state the following

Theorem 2.1. *Let $\widetilde{y}_{n+1}$ be the approximation provided by a one-step method of order p and leading term of the local error $h^{p+1}\widetilde{\varphi}_{p+1}$. Let us suppose that we are computing a projected solution using a direction $w_n = h^{q+1}\psi_{q+1}+\mathcal{O}(h^{q+2})$ with $q < p$ and $\nabla G(y_n)\cdot\psi_{q+1} \neq 0$. Then, the projected solution $y_{n+1} = \widetilde{y}_{n+1} + \lambda_n w_n$, with λ_n such that $G(y_{n+1}) = G(y_n)$ has order $\geq p+1$ if and only if the vectors $\widetilde{\varphi}_{p+1}$ and ψ_{q+1} are parallel.*

Proof. By Theorem 4.1 of [7], there exists a unique λ_n such that $G(y_{n+1}) = G(\widetilde{y}_{n+1} + \lambda_n w_n) = G(y_n)$ with $\lambda_n = \sigma_{p-q}h^{p-q} + \sigma_{p-q+1}h^{p-q+1} + \mathcal{O}(h^{p-q+2})$. Therefore, the expansion of the local error of the projected solution will be given by

$$\begin{aligned} y(t_n+h;t_n,y_n) - \; & y_{n+1} = y(t_n+h;t_n,y_n) - (\widetilde{y}_{n+1} + \lambda_n w_n) \\ = \; & h^{p+1}\widetilde{\varphi}_{p+1} + h^{p+2}\widetilde{\varphi}_{p+2} + \mathcal{O}(h^{p+3}) \\ & -(\sigma_{p-q}h^{p-q} + \sigma_{p-q+1}h^{p-q+1})(h^{q+1}\psi_{q+1} + h^{q+2}\psi_{q+2}) + \mathcal{O}(h^{p+3}) \\ = \; & h^{p+1}(\widetilde{\varphi}_{p+1} - \sigma_{p-q}\psi_{q+1}) \\ & +h^{p+2}(\widetilde{\varphi}_{p+2} - \sigma_{p-q}\psi_{q+2} - \sigma_{p-q+1}\psi_{q+1}) + \mathcal{O}(h^{p+3}). \end{aligned}$$

On the other hand

$$\begin{aligned} 0 = \; & G(y_{n+1}) - G(y_n) = G\big(y(t_{n+1};t_n,y_n) + y_{n+1} - y(t_{n+1};t_n,y_n)\big) - G(y_n) \\ = \; & G(y(t_{n+1};t_n,y_n)) - G(y_n) - h^{p+1}\nabla G(y(t_{n+1};t_n,y_n)).(\widetilde{\varphi}_{p+1} - \sigma_{p-q}\psi_{q+1}) + \mathcal{O}(h^{p+2}) \\ = \; & -h^{p+1}\nabla G(y_n).(\widetilde{\varphi}_{p+1} - \sigma_{p-q}\psi_{q+1}) + \mathcal{O}(h^{p+2}), \end{aligned}$$

which means that

$$\nabla G(y_n).(\widetilde{\varphi}_{p+1} - \sigma_{p-q}\psi_{q+1}) = 0.$$

Denoting by $u = \nabla G(y_n)$, we can decompose $\widetilde{\varphi}_{p+1} = \alpha u + \widetilde{v}$ and $\psi_{q+1} = \beta u + v$ with $\widetilde{v}$ and v vectors orthogonal to u and $\beta \neq 0$. Therefore

$$\sigma_{p-q} = \alpha/\beta,$$

and $\widetilde{\varphi}_{p+1} - \sigma_{p-q}\psi_{q+1} = \widetilde{v} - (\alpha/\beta)v$. Now, since $\widetilde{\varphi}_{p+1}$ and ψ_{q+1} are parallel if and only if $\widetilde{v} = (\alpha/\beta)v = \sigma_{p-q}v$, $\widetilde{\varphi}_{p+1} - \sigma_{p-q}\psi_{q+1}$ vanishes if and only if $\widetilde{\varphi}_{p+1}$ and ψ_{q+1} are parallel. □

Notice that the above result depends neither on how the direction w_n is obtained nor on the order q ($q < p$). Then if we had an asymptotically correct estimation of the direction of the leading error term $\widetilde{\varphi}_{p+1}$ of the solution of order p, we could construct a projection method that provides a higher order approximation.

On the other hand, the vectors must be parallel for all n, that is at any step. This makes this property very dependent on the problem being integrated and it is unlikely to find projectors that gain an order. A simple

example of such a problem is the harmonic oscillator $z'' = -\omega^2 z$. It can be expressed as the linear system of first order

$$y' = \omega J y, \quad J = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

and it has the first integral $G(y) = y_1^2 + y_2^2 = y^T y$, with $y = (y_1, y_2)^T$.

For this problem it is easy to see that all the elementary differentials vanish except those corresponding to the one-branch trees, and they have the form $f' \cdots f' f(y)$. Therefore, for any Runge–Kutta method of order p the vector $\widetilde{\varphi}_{p+1}(y_n)$ will be parallel to the vector y_n if p is odd, and parallel to the vector Jy_n if p is even.

Then we have the following

Corollary 2.1. *Let us consider a Runge–Kutta method of odd order p. For the harmonic oscillator we have:*

- *The order of the simplified orthogonal projection is at least $p+1$.*
- *The order of the RK projected solution is at least $p+1$ if and only if the embedded method has order q odd ($q < p$).*

Proof. The gradient of the invariant is $\nabla G(y) = 2y^T$. Therefore, if the non-projected solution is $\widetilde{y}_{n+1}$, the direction vector for the simplified orthogonal projection is $\nabla G(\widetilde{y}_{n+1})^T = 2y_n + \mathcal{O}(h)$ which is asymptotically parallel to the vector y_n. Since p is odd, the vector $\widetilde{\varphi}_{p+1}$ is parallel to the vector y_n, and therefore the projected solution will have order at least $p+1$.

For a Runge–Kutta projection, based on a embedded formula, the vector w_n will be parallel to y_n if q is odd and in this case the order of the projected solution will be at least $p+1$.

However, if q is even, w_n will be asymptotically parallel to Jy_n, that is, orthogonal to y_n, which is the direction of $\widetilde{\varphi}_{p+1}$, and according to Theorem 2.1, the order of the projected solution is at most p. Notice that in this case, the order p can not be guaranteed. □

From Theorem 2.1 it is also seen that the leading term of the local error of the projected approximation of order p depends only on the vector $\widetilde{\varphi}_{p+1}$ and the direction determined by ψ_{q+1} and not on the magnitude of this last vector. Notice that the vector coefficient of the leading term of the local error is $(\widetilde{\varphi}_{p+1} - \sigma_{p-q}\psi_{q+1})$. But if we scale the vector ψ_{q+1} by a factor $k \neq 0$, then the scalar σ_{p-q} will be scaled by $1/k$. Therefore, all the embedded methods that give the same direction ψ_{q+1} provide projected approximations with the same leading term. This allows us to give the following

Theorem 2.2. *All the RK projection methods based on a RK formula of order p and an embedded formula of order $q = 1$, have the same leading term of the local error, assumed that it has order p, but not $(p+1)$. If the*

embedded formula has order $q > 1$, there is at most an $(r-1)$-parameter family of projected methods with different leading terms of the local error, r being the number of elementary differentials of order $(q+1)$.

Proof. If the embedded method has order $q = 1$,

$$\widehat{y}_{n+1} - \widetilde{y}_{n+1} = -\frac{h^2}{2}(1 - 2\,\widehat{b}^T c) f' f(y_n) + \mathcal{O}(h^3).$$

Then, the vector ψ_2 is always parallel to the elementary differential $f'f$ and the leading term of the local error of the projected solution is always the same.

If the embedded method has order $q = 2 < p$,

$$\widehat{y}_{n+1} - \widetilde{y}_{n+1} = -\frac{h^3}{6}\big[(1 - 3\,\widehat{b}^T c^2) f''(f,f)(y_n) + (1 - 6\,\widehat{b}^T A c) f' f' f(y_n)\big] + \mathcal{O}(h^4).$$

In this case, the possible directions for the vector ψ_3 are those obtained by a linear combination of the two elementary differentials of order 3, that are of the form $f'f'f(y_n)$ or $f''(f,f)(y_n) + \gamma f'f'f(y_n)$, with γ as free parameter. The generalization for order $q < p$ is straightforward. □

3 RK projection methods for oscillatory problems

The harmonic oscillator is the simplest test problem when solving oscillatory problems. From Corollary 2.1 we know that for RK projected methods applied to these linear equations, the order of the approximation is $p + 1$ if p and q are odd. Next, we will see that the order can be higher if the embedded formula is chosen appropriately.

For this class of problems it is usual to decompose the error of the numerical methods in two parts: the dissipation and the dispersion errors. Thus, we consider the complex scalar test equation

$$y' = i\omega y, \quad y(0) = y_0 \in \mathbb{C},$$

where $i = \sqrt{-1}$, and ω is a real constant. The solution is the complex exponential $y(t) = y_0 e^{i\omega t}$, and it has the invariant $G(y) = |y|^2$.

The numerical solution provided by a RK method (A, b) satisfies

$$y_{n+1} = R(i\nu) y_n, \tag{6}$$

where $\nu = h\omega$, and

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

with $e = (1, \ldots, 1)^T \in \mathbb{R}^s$. Thus, a comparison of (6) with $y(t_{n+1}) = e^{i\nu} y(t_n)$, leads to decompose the error into the dispersion and dissipation errors, defined [6, 17, 18] as the errors in phase and modulus respectively

$$\phi(\nu) := \nu - \arg(R(i\nu)) = \nu - \arctan\left(\frac{\mathrm{Im}(R(i\nu))}{\mathrm{Re}(R(i\nu))}\right),$$
$$d(\nu) := 1 - |R(i\nu)|.$$

In addition, if the dispersion (resp. dissipation) error is $O(\nu^{r+1})$, the RK method is said to be dispersive (resp. dissipative) of order r. In the following, to simplify calculations, we will refer to the dissipation error as $d(\nu) := 1 - |R(i\nu)|^2$. This last definition is equivalent to the previous one in the sense that it leads to the same dissipation order.

Let us suppose that we are using a RK formula of order p with stability function $\widetilde{R}(z)$ as the main integrator and another scheme of order $q < p$ with stability function $\widehat{R}(z)$. The projected method will have as stability function $R(z) = (1 - \lambda_n)\widetilde{R}(z) + \lambda_n \widehat{R}(z)$, where the coefficient λ_n is computed so that

$$|R(i\nu)|^2 = |(1 - \lambda_n)\widetilde{R}(i\nu) + \lambda_n \widehat{R}(i\nu)|^2 = 1, \tag{7}$$

This implies that for any direction w, the projected method is zero dissipative, and the error reduces to the dispersion error. Next, we will study the order of projection methods based on the third order scheme of Bogacki and Shampine.

3.1 Projection method based on Bogacki–Shampine RK method

Let us take as basic method the 3-stage RK method of order 3 derived by Bogacki and Shampine in [1], with coefficients $(A, \widetilde{b}^T)$. This is a well known explicit RK formula used in the Matlab ODE suite package [24]. To get the projected solution we have to choose the coefficients $\widehat{b}_i$, $i = 1, 2$, of $\widehat{y}_{n+1}$, consistent RK method embedded to $\widetilde{y}_{n+1}$. Thus, the coefficients of this embedded RK pair are given by the Butcher's tableau:

$$\begin{array}{c|c} c & A \\ \hline & \widetilde{b}^T \\ \hline & \widehat{b}^T \end{array} \quad = \quad \begin{array}{c|ccc} 0 & & & \\ 1/2 & 1/2 & & \\ 3/4 & 0 & 3/4 & \\ \hline & 2/9 & 1/3 & 4/9 \\ \hline & \widehat{b}_1 & \widehat{b}_2 & 1 - \widehat{b}_1 - \widehat{b}_2 \end{array} \tag{8}$$

For the Bogacki–Shampine method,

$$\begin{aligned}\widetilde{R}(z) &= 1+z+\frac{z^2}{2}+\frac{z^3}{6}, \quad (z=i\nu),\\ \widetilde{\phi}(\nu) &= -\frac{1}{30}\nu^5+\frac{1}{252}\nu^7+O(\nu^9),\\ \widetilde{d}(\nu) &= \frac{1}{12}\nu^4-\frac{1}{36}\nu^6,\end{aligned}$$

so it is dispersive of order 4 and dissipative of order 3.

3.1.1 Embedded formulas of order one

The 3-stage method $\widehat{y}_{n+1}$ embedded to $\widetilde{y}_{n+1}$, has strict order one if $-1+3\widehat{b}_1+\widehat{b}_2\neq 0$. In this case

$$\begin{aligned}\widehat{R}(z) &= 1+z+\frac{1}{4}(3-3\widehat{b}_1-\widehat{b}_2)z^2+\frac{3}{8}(1-\widehat{b}_1-\widehat{b}_2)z^3, \quad (z=i\nu),\\ \widehat{\phi}(\nu) &= \frac{1}{24}(-1+9\widehat{b}_1-3\widehat{b}_2)\nu^3+O(\nu^5),\\ \widehat{d}(\nu) &= \frac{1}{2}(1-3\widehat{b}_1-\widehat{b}_2)\nu^2+\frac{1}{16}(3+6\widehat{b}_1-9\widehat{b}_1^2-6\widehat{b}_2-6\widehat{b}_1\widehat{b}_2-\widehat{b}_2^2)\nu^4-\\ &\quad \frac{9}{64}(1-\widehat{b}_1-\widehat{b}_2)^2\nu^6.\end{aligned}$$

Solving (7) in λ_n and taking the solution that gives rise to the lowest dispersion error for the projected approximation y_{n+1} given in (5), we obtain:

$$\phi(\nu)=\frac{19-27\widehat{b}_1-39\widehat{b}_2}{720(-1+3\widehat{b}_1+\widehat{b}_2)}\nu^5+O(\nu^7).$$

So, since $-1+3\widehat{b}_1+\widehat{b}_2\neq 0$, the approximation y_{n+1} is dispersive of order four, and this order is six if and only if

$$19-27\widehat{b}_1-39\widehat{b}_2=0,$$

and in this case,

$$\phi(\nu)=\frac{1}{12600}\nu^7+O(\nu^9).$$

We have therefore a one-parameter family of embedded methods of order one that give a projected zero-dissipative solution of dispersion order six, and consequently of order six for the linear test equation. Nevertheless, according to Theorem 2.2, since $q=1$, all these methods give projected solutions with the same leading local error term for general problems. Consequently, all of

them are asymptotically equivalent. Even more, for these embedded methods the expansion of $w_n = \widehat{y}_{n+1} - \widetilde{y}_{n+1}$ is

$$\begin{aligned}\widehat{y}_{n+1} - \widetilde{y}_{n+1} &= 5(2 - 9\widehat{b}_1)h^2 f'f(y_n)/78 + \\ &\quad (2 - 9\widehat{b}_1)h^3\big[3f''(f,f)(y_n) + 2f'f'f(y_n)\big]/156 + \mathcal{O}(h^4)\end{aligned}$$

and not only the direction of ψ_2 is the same for all the methods of the family, but also the direction of the vector ψ_3.

3.1.2 Embedded formulas of order two

The 3-stage method $\widehat{y}_{n+1}$ embedded to $\widetilde{y}_{n+1}$, has order two if $-1+3\widehat{b}_1+\widehat{b}_2 = 0$ and $\widehat{b}_1 \neq 2/9$. It results

$$\begin{aligned}\widehat{R}(z) &= 1 + z + \frac{1}{2}z^2 + \frac{3}{4}\widehat{b}_1 z^3, \quad (z = i\nu), \\ \widehat{\phi}(\nu) &= \frac{1}{12}(9\widehat{b}_1 - 2)\nu^3 + O(\nu^5), \\ \widehat{d}(\nu) &= \frac{1}{4}(6\widehat{b}_1 - 1)\nu^4 - \frac{9}{16}\widehat{b}_1^2\nu^6\end{aligned}$$

Solving (7) in λ_n and taking the solution that gives rise to the lowest dispersion error for the projected approximation y_{n+1}, we obtain:

$$\phi(\nu) = -\frac{1}{24}\nu^3 + O(\nu^5).$$

So, the value y_{n+1} is dispersive of order 2. Note that in this case, the gradient of the invariant $\nabla G(y_n)$, and the vector w_n are orthogonal, and this is the reason why the projected approximation loses an order with respect to the underlying RK scheme for the harmonic oscillator.

We have therefore a one-parameter family of embedded methods of order two that give a projected solution of order three in general, but of order two for the linear test equation. In addition, for these embedded methods the expansion of $w_n = \widehat{y}_{n+1} - \widetilde{y}_{n+1}$ is

$$\widehat{y}_{n+1} - \widetilde{y}_{n+1} = (9\widehat{b}_1 - 2)\big[f''(f,f)(y_n) + 4f'f'f(y_n)\big]h^3/48 + \mathcal{O}(h^4)$$

and the direction of ψ_3 is the same for all the methods of the family. According to Theorem 2.1, all these methods give the same leading term of the local error for general problems. All of them are asymptotically equivalent.

4 RK projection methods for general problems

For general problems we will consider, for small enough stepsize h, the following two criteria:

(C_1) If $\widetilde{y}_{n+1}$ advances the phase with respect to $y(t_{n+1}; t_n, y_n)$, then $\widehat{y}_{n+1}$ must delay it, and conversely.

(C_2) $g(\widetilde{y}_{n+1})$ and $g(\widehat{y}_{n+1})$ have opposite signs.

The idea is on one side to ensure the existence of the scalar parameter λ_n in each step, since by (C_2) there exists λ_n, $0 < \lambda_n < 1$, such that $g(y_{n+1}) = 0$. On the other hand, (C_1) together with (C_2) may contribute to reduce the global errors $|y(t_n) - y_n|$ of the projected method.

We assume for $\widehat{y}_{n+1}$ the consistence condition, so $\widehat{b} = (\widehat{b}_1, \widehat{b}_2, \ldots, \widehat{b}_{s-1}, 1 - \sum_{i=1}^{s-1} \widehat{b}_i)^T$, and we have $s-1$ free parameters.

Note that for the linear test equation condition (C_1) reduces to

$$\widetilde{\phi}(\nu)\, \widehat{\phi}(\nu) < 0, \tag{9}$$

for $\nu \in \mathbb{R}$ small enough, whereas condition (C_2) means

$$\widetilde{d}(\nu)\, \widehat{d}(\nu) < 0. \tag{10}$$

In relation with (C_2), our first task is how to approximate $g(\widehat{y}_{n+1})$. If $\widetilde{y}_{n+1}$ and $\widehat{y}_{n+1}$ have orders p and q, respectively, $p > q$, we can write

$$\begin{aligned} g(\widehat{y}_{n+1}) &= g(\widetilde{y}_{n+1} + (\widehat{y}_{n+1} - \widetilde{y}_{n+1})) \\ &= g(\widetilde{y}_{n+1}) + \nabla g(\widetilde{y}_{n+1}) \cdot (\widehat{y}_{n+1} - \widetilde{y}_{n+1}) + O(h^{2(q+1)}). \end{aligned}$$

Since $g(\widetilde{y}_{n+1}) = O(h^{p+1})$ and $\nabla g(\widetilde{y}_{n+1}) \cdot (\widehat{y}_{n+1} - \widetilde{y}_{n+1}) = O(h^{q+1})$, the dominant term in $g(\widehat{y}_{n+1})$ is

$$\nabla g(\widetilde{y}_{n+1}) \cdot (\widehat{y}_{n+1} - \widetilde{y}_{n+1}) = h \sum_{i=1}^{s} (\widehat{b}_i - \widetilde{b}_i) \nabla g(\widetilde{y}_{n+1}) \cdot g_i = h \sum_{i=1}^{s} (\widehat{b}_i - \widetilde{b}_i) k_i,$$

where $g_i = f(y_n + h \sum_{j=1}^{i-1} a_{ij} g_j)$, $i = 1, 2, \ldots, s$, are the stages of the embedded pair, and we have introduced the scalars

$$k_i = \nabla g(\widetilde{y}_{n+1}) \cdot g_i, \;\; i = 1, \ldots, s. \tag{11}$$

Since $g(\widetilde{y}_{n+1})$ is known, we will approximate

$$g(\widehat{y}_{n+1}) \approx \widehat{g}(\widehat{y}_{n+1}) := g(\widetilde{y}_{n+1}) + h \sum_{i=1}^{s} (\widehat{b}_i - \widetilde{b}_i) k_i,$$

and we will choose in each step appropriate coefficients $\widehat{b}_i$ so that

$$g(\widetilde{y}_{n+1})\,\widehat{g}(\widehat{y}_{n+1}) < 0. \tag{12}$$

Obviously, (C_2) can be applied if $g(\widetilde{y}_{n+1}) \neq 0$. If $g(\widetilde{y}_{n+1}) = 0$, we do not need to project and we will simply take $y_{n+1} = \widetilde{y}_{n+1}$. In addition, because of the consistency of both methods, we have

$$\sum_{i=1}^{s}(\widehat{b}_i - \widetilde{b}_i)k_i = \sum_{i=1}^{s-1}(k_i - k_s)\widehat{b}_i + \sum_{i=1}^{s-1}(k_s - k_i)\widetilde{b}_i,$$

and so, if $k_i = k_j\ \forall i \neq j$, which is unlikely, then (C_2) represented by (12) does not make any sense since, in this case, $\widehat{g}(\widehat{y}_{n+1}) = g(\widetilde{y}_{n+1})$.

Regarding condition (C_1), we must fix the meaning of "advance the phase". A possibility is to consider the angle formed by the vectors y_n and y_{n+1}. We can say that the numerical solution advances the phase if it is greater than the angle formed by y_n and the local solution $y(t_n + h; t_n, y_n)$, that is,

$$\frac{y_n^T y_{n+1}}{\|y_n\|\|y_{n+1}\|} > \frac{y_n^T y(t_n + h; t_n, y_n)}{\|y_n\|\|y(t_n + h; t_n, y_n)\|}.$$

Assuming that $\|y(t_{n+1})\| \simeq \|y_{n+1}\|$, this condition can be approximated by

$$y_n^T\,(y_{n+1} - y(t_n + h; t_n, y_n)) > 0$$

If we had a good estimation of the direction of the local error $y_{n+1} - y(t_n + h; t_n, y_n)$, we could estimate the above condition, but in general, this is not possible. In the following, we will consider the phase determined by the linear test equation.

4.1 Projection method based on Bogacki–Shampine RK method

In this section we will study how to get appropriate embedded RK methods $(A, \widehat{b})$ with coefficients given in (8) according to the ideas exposed before. For that purpose, in what follows we will assume that we are looking for approximations $\widehat{y}_{n+1}$ of order one, which are dispersive of order two and dissipative of order one, according to section 3.1.1.

The next result establishes conditions to get approximations $\widehat{y}_{n+1}$ satisfying (C_1) for the harmonic oscillator, (C_2) for general problems and such that the projected method is zero dissipative with the highest dispersion order possible.

Theorem 4.1. *There exist embedded RK methods $(A, \widehat{b}^T)$ given in* (8)*, satisfying* (9)*,* (10) *and* (12)*, such that the projected solution y_{n+1} obtained according to* (5) *has order 6 for the harmonic oscillator, if and only if*

$$\operatorname{sign} g(\widetilde{y}_{n+1}) = -\operatorname{sign}(13k_1 - 9k_2 - 4k_3),$$

with k_i, $i = 1, 2, 3$ *given in* (11). *In such a case, the coefficients of the embedded methods must satisfy*

$$\widehat{b}_1 > \frac{2}{9} - \frac{13g(\widetilde{y}_{n+1})}{h(13k_1 - 9k_2 - 4k_3)}, \quad \widehat{b}_2 = \frac{19}{39} - \frac{9}{13}\widehat{b}_1.$$

Proof. Let $\widehat{y}_{n+1}$ be the approximation provided by the embedded method. From section 3.1.1 we know that, if $\widehat{y}_{n+1}$ has order one, i.e., $-1+3\widehat{b}_1+\widehat{b}_2 \neq 0$, the projected solution is dispersive of order 6 if and only if

$$19 - 27\widehat{b}_1 - 39\widehat{b}_2 = 0,$$

that is,

$$\widehat{b}_2 = \frac{19}{39} - \frac{9}{13}\widehat{b}_1, \tag{13}$$

and, under this condition, (9)–(10) are equivalent to

$$\widehat{b}_1 > \frac{2}{9}. \tag{14}$$

After substituting the value for $\widehat{b}_2$ obtained in (13) and the coefficients $\widetilde{b}_i$, $i = 1, 2, 3$, into $\widehat{g}(\widehat{y}_{n+1})$, it is obtained that

$$\widehat{g}(\widehat{y}_{n+1}) = g(\widetilde{y}_{n+1}) + \frac{h}{117}(9\widehat{b}_1 - 2)(13k_1 - 9k_2 - 4k_3),$$

where $9\widehat{b}_1 - 2 > 0$. Then, if either $13k_1 - 9k_2 - 4k_3 = 0$ or else sign $g(\widetilde{y}_{n+1}) =$ sign $(13k_1 - 9k_2 - 4k_3)$, it is clear that (12) is not satisfied. Otherwise, i.e. if sign $g(\widetilde{y}_{n+1}) = -$sign $(13k_1 - 9k_2 - 4k_3)$, the values for $\widehat{b}_1$ satisfying (12) are given by

$$\widehat{b}_1 > \frac{2}{9} - \frac{13g(\widetilde{y}_{n+1})}{h(13k_1 - 9k_2 - 4k_3)} > \frac{2}{9},$$

which completes the proof. □

In particular, if $k_1 = k_2 = k_3$ and so, condition (12) can not be satisfied, we will choose the coefficients $\widehat{b}_1$, $\widehat{b}_2$ according to (13) and (14), i.e. satisfying the rest of conditions imposed in the previous theorem.

Theorem 4.1 shows that it is not always possible to find $\widehat{y}_{n+1}$ satisfying all the conditions imposed there. Next, we are going to weaken them so that the attainment of appropriate approximations $\widehat{y}_{n+1}$ is always assured. Thus, concerning the criterium (C_2), and more specifically, the condition (12), we have

Theorem 4.2. *The RK method* $(A, \widehat{b}^T)$ *whose coefficients are given in* (8)*, satisfies* (12) *if and only if one of these four situations holds:*

a)

$$\begin{cases} \operatorname{sign} g(\widetilde{y}_{n+1}) = \operatorname{sign}(k_2 - k_3), \\ \widehat{b}_1 \ \textit{arbitrary}, \ \widehat{b}_2 < \alpha(\widehat{b}_1), \end{cases}$$

b)

$$\begin{cases} \operatorname{sign} g(\widetilde{y}_{n+1}) = -\operatorname{sign}(k_2 - k_3), \\ \widehat{b}_1 \ \textit{arbitrary}, \ \widehat{b}_2 > \alpha(\widehat{b}_1), \end{cases}$$

c)

$$\begin{cases} k_2 = k_3, \ \operatorname{sign} g(\widetilde{y}_{n+1}) = \operatorname{sign}(k_1 - k_3), \\ \widehat{b}_1 < \beta, \ \widehat{b}_2 \ \textit{arbitrary}, \end{cases}$$

d)

$$\begin{cases} k_2 = k_3, \ \operatorname{sign} g(\widetilde{y}_{n+1}) = -\operatorname{sign}(k_1 - k_3), \\ \widehat{b}_1 > \beta, \ \widehat{b}_2 \ \textit{arbitrary}, \end{cases}$$

where $k_i, \ i = 1, 2, 3$ *are given in* (11)*, and*

$$\begin{aligned} &\alpha(\widehat{b}_1) = \frac{k_3 - k_1}{k_2 - k_3}\widehat{b}_1 + \frac{2k_1 + 3k_2 - 5k_3}{9(k_2 - k_3)} - \frac{g(\widetilde{y}_{n+1})}{h(k_2 - k_3)}, \quad (k_2 \neq k_3), \\ &\beta = \frac{2}{9} - \frac{g(\widetilde{y}_{n+1})}{h(k_1 - k_3)}, \quad (k_1 \neq k_3). \end{aligned} \tag{15}$$

Proof. The substitution of the coefficients $\widetilde{b}_i$, $i = 1, 2, 3$, according to (8) into $\widehat{g}(\widehat{y}_{n+1})$ gives rise to

$$\widehat{g}(\widehat{y}_{n+1}) = g(\widetilde{y}_{n+1}) + h\Big[(k_1 - k_3)\widehat{b}_1 + (k_2 - k_3)\widehat{b}_2 - \frac{2k_1 + 3k_2 - 5k_3}{9}\Big].$$

If $k_1 = k_2 = k_3$, then clearly (12) is not satisfied. Otherwise, $\widehat{g}(\widehat{y}_{n+1}) < 0$ is equivalent to

$$\begin{cases} \widehat{b}_2 < \alpha(\widehat{b}_1), & \text{if } k_2 - k_3 > 0; \\ \widehat{b}_2 > \alpha(\widehat{b}_1), & \text{if } k_2 - k_3 < 0; \\ \widehat{b}_1 < \beta, & \text{if } k_2 = k_3 \ \& \ k_1 - k_3 > 0; \\ \widehat{b}_1 > \beta, & \text{if } k_2 = k_3 \ \& \ k_1 - k_3 < 0; \end{cases}$$

and $\widehat{g}(\widehat{y}_{n+1}) > 0$ is equivalent to

$$\begin{cases} \widehat{b}_2 < \alpha(\widehat{b}_1), & \text{if } k_2 - k_3 < 0; \\ \widehat{b}_2 > \alpha(\widehat{b}_1), & \text{if } k_2 - k_3 > 0; \\ \widehat{b}_1 < \beta, & \text{if } k_2 = k_3 \ \& \ k_1 - k_3 < 0; \\ \widehat{b}_1 > \beta, & \text{if } k_2 = k_3 \ \& \ k_1 - k_3 > 0; \end{cases}$$

The proof follows from the above inequalities. □

Therefore, for the embedded RK pair constructed from the Bogacki–Shampine method, if (C_2) can be applied, i.e. if $g(\widetilde{y}_{n+1}) \neq 0$, there always exist coefficients $\widehat{b}_1$, $\widehat{b}_2$ satisfying (12) unless $k_1 = k_2 = k_3$, in which case (12) does not make any sense as commented before.

Next, we will study when the combination of the criterium (C_1) for the harmonic oscillator and (C_2) is possible.

Theorem 4.3. *The RK method $\widehat{y}_{n+1}$ with coefficients $(A, \widehat{b}^T)$ given in* (8) *satisfies* (9) *and also* (12) *if and only if one of these four situations happens:*

a)

$$\begin{cases} \operatorname{sign} g(\widetilde{y}_{n+1}) = \operatorname{sign}(k_2 - k_3), \\ \widehat{b}_1 \ \textit{arbitrary}, \ \widehat{b}_2 < \min\Big\{ -\dfrac{1}{3} + 3\widehat{b}_1, \alpha(\widehat{b}_1) \Big\}. \end{cases}$$

b)

$$\begin{cases} \operatorname{sign} g(\widetilde{y}_{n+1}) = -\operatorname{sign}(k_2 - k_3), \\ \widehat{b}_1 \begin{cases} < \gamma, & \textit{if} \operatorname{sign} \ (k_2 - k_3) = -\operatorname{sign} \ (k_1 + 3k_2 - 4k_3), \\ > \gamma, & \textit{if} \operatorname{sign} \ (k_2 - k_3) = \operatorname{sign} \ (k_1 + 3k_2 - 4k_3), \end{cases} \\ \widehat{b}_2 \ \ni \ \alpha(\widehat{b}_1) < \widehat{b}_2 < -\dfrac{1}{3} + 3\widehat{b}_1. \end{cases}$$

c)

$$\begin{cases} k_2 = k_3, \quad \operatorname{sign} g(\widetilde{y}_{n+1}) = \operatorname{sign}(k_1 - k_3), \\ \widehat{b}_2 < -\dfrac{1}{3} + 3\beta, \ \widehat{b}_1 \ \ni \ \dfrac{1}{9} + \dfrac{\widehat{b}_2}{3} < \widehat{b}_1 < \beta. \end{cases}$$

d)

$$\begin{cases} k_2 = k_3, \ \operatorname{sign} g(\widetilde{y}_{n+1}) = -\operatorname{sign}(k_1 - k_3), \\ \widehat{b}_2 \ \textit{arbitrary}, \ \widehat{b}_1 > \max\{\beta, \dfrac{1}{9} + \dfrac{\widehat{b}_2}{3}\}. \end{cases}$$

where $k_i,\ i = 1, 2, 3$ *are given in* (11), $\alpha(\widehat{b}_1)$ *and* β *in* (15) *and*

$$\gamma = \frac{2}{9} - \frac{g(\widetilde{y}_{n+1})}{h(k_1 + 3k_2 - 4k_3)}, \quad (k_1 + 3k_2 - 4k_3 \neq 0).$$

Proof. Taking into account the principal term of the dispersion error $\widetilde{\phi}(\nu)$ and $\widehat{\phi}(\nu)$, condition (9) is equivalent to

$$-1 + 9\widehat{b}_1 - 3\widehat{b}_2 > 0. \tag{16}$$

Item a) is a direct consequence of (16) and item a) in Theorem 4.2. Condition (16) together with item b) in Theorem 4.2 lead to coefficients $\widehat{b}_1$, $\widehat{b}_2$ satisfying

$$\alpha(\widehat{b}_1) < \widehat{b}_2 < -\frac{1}{3} + 3\widehat{b}_1.$$

In particular, $\widehat{b}_1$ must be such that $\alpha(\widehat{b}_1) < -\frac{1}{3} + 3\widehat{b}_1$, which is equivalent to

$$\frac{k_1 + 3k_2 - 4k_3}{k_2 - k_3}\,\widehat{b}_1 > \frac{2(k_1 + 3k_2 - 4k_3)}{9(k_2 - k_3)} - \frac{g(\widetilde{y}_{n+1})}{h(k_2 - k_3)}.$$

Here, $k_1 + 3k_2 - 4k_3 \neq 0$, since otherwise the above expression would be equivalent to $\frac{g(\widetilde{y}_{n+1})}{k_2 - k_3} > 0$, which is not true in this case. So, from that expression we obtain item b).

Condition (16) and the restriction for $\widehat{b}_1$ in item c) in Theorem 4.2 conduce to

$$\frac{1}{9} + \frac{\widehat{b}_2}{3} < \widehat{b}_1 < \beta,$$

and so we have the actual item c). Finally, it is clear that (16) together with item d) in Theorem 4.2 give rise to item d) in the present Theorem. □

Therefore, if (C_2) can be applied (i.e. $g(\widetilde{y}_{n+1}) \neq 0$) and its application from (12) makes sense (i.e. $k_1 = k_2 = k_3$ does not happen), we have proved that there always exist coefficients $\widehat{b}_1$, $\widehat{b}_2$ for the RK method $\widehat{y}_{n+1}$ given in (8) satisfying both criteria (C_1) and (C_2), unless in this case:

$$\text{sign } g(\widetilde{y}_{n+1}) = -\text{sign } (k_2 - k_3),\ (k_2 \neq k_3), \quad k_1 + 3k_2 - 4k_3 = 0.$$

If these two conditions are satisfied, which is quite unlikely, we wonder if there exist coefficients $\widehat{b}_1, \widehat{b}_2$ of $\widehat{y}_{n+1}$, dispersive of order 2, satisfying

$$\widetilde{\phi}(\nu)\widehat{\phi}(\nu) > 0,\ (\nu \to 0), \tag{17}$$
$$g(\widetilde{y}_{n+1})\widehat{g}(\widehat{y}_{n+1}) > 0. \tag{18}$$

From (17) we obtain $\widehat{b}_2 > -\frac{1}{3} + 3\widehat{b}_1$. After substituting $k_1 = 4k_3 - 3k_2$ into $\widehat{g}(\widehat{y}_{n+1})$, we have

$$\widehat{g}(\widehat{y}_{n+1}) = g(\widetilde{y}_{n+1}) + h(k_2 - k_3)\Big(- 3\widehat{b}_1 + \widehat{b}_2 + \frac{1}{3}\Big),$$

and since sign $g(\widetilde{y}_{n+1}) = -$sign $(k_2 - k_3)$, condition (18) conduces to

$$\widehat{b}_2 < 3\widehat{b}_1 - \frac{1}{3} - \frac{g(\widetilde{y}_{n+1})}{h(k_2 - k_3)}.$$

Therefore, $\widehat{b}_2$ must be such that

$$-\frac{1}{3} + 3\widehat{b}_1 < \widehat{b}_2 < 3\widehat{b}_1 - \frac{1}{3} - \frac{g(\widetilde{y}_{n+1})}{h(k_2 - k_3)}.$$

There is no restriction now for $\widehat{b}_1$.

In practice, if a parameter is lower (resp. greater) than a given value, e.g. $\widehat{b}_1 < \gamma$ (resp. $\widehat{b}_1 > \gamma$), we will take $\widehat{b}_1 = \gamma - \epsilon$ (resp. $\widehat{b}_1 = \gamma + \epsilon$) for some $\epsilon > 0$. In addition, if a parameter must be between two given values, we will take it as the average of both values. Thus, taking into account the previous results, we are going to give an algorithm to obtain in each step an appropriate approximation $\widehat{y}_{n+1}$ embedded to the Bogacki–Shampine method $\widetilde{y}_{n+1}$. The algorithm procedes as follows:

Algorithm to obtain $\widehat{b}_1$ and $\widehat{b}_2$ in (8)

1. If $g(\widetilde{y}_{n+1}) = 0$, then we do not project, and we take $y_{n+1} = \widetilde{y}_{n+1}$.

2. If $k_1 = k_2 = k_3$, then:

$$\widehat{b}_1 = \frac{2}{9} + \epsilon, \quad \widehat{b}_2 = \frac{19}{39} - \frac{9}{13}\widehat{b}_1.$$

3. Else if sign $g(\widetilde{y}_{n+1}) = -$sign$(13k_1 - 9k_2 - 4k_3)$, then:

$$\widehat{b}_1 = \frac{2}{9} - \frac{13g(\widetilde{y}_{n+1})}{h(13k_1 - 9k_2 - 4k_3)} + \epsilon, \quad \widehat{b}_2 = \frac{19}{39} - \frac{9}{13}\widehat{b}_1.$$

   We have in this case a projection method that satisfies (9), (10) and (12), and it has order 6 for the harmonic oscillator.

4. Else if sign $g(\widetilde{y}_{n+1}) =$ sign $(k_2 - k_3)$, then:

$$\widehat{b}_1 \text{ arbitrary}, \quad \widehat{b}_2 = \min\Big\{ - \frac{1}{3} + 3\widehat{b}_1, \alpha(\widehat{b}_1)\Big\} - \epsilon.$$

5. Else if sign $g(\widetilde{y}_{n+1}) = -\text{sign}(k_2 - k_3) = \text{sign}\ (k_1 + 3k_2 - 4k_3)$, then:

$$\widehat{b}_1 = \gamma - \epsilon, \quad \widehat{b}_2 = \frac{\alpha(\widehat{b}_1)}{2} - \frac{1}{6} + \frac{3}{2}\widehat{b}_1.$$

6. Else if sign $g(\widetilde{y}_{n+1}) = -\text{sign}(k_2 - k_3) = -\text{sign}\ (k_1 + 3k_2 - 4k_3)$, then:

$$\widehat{b}_1 = \gamma + \epsilon, \quad \widehat{b}_2 = \frac{\alpha(\widehat{b}_1)}{2} - \frac{1}{6} + \frac{3}{2}\widehat{b}_1.$$

7. Else if $(k_2 = k_3)$ and sign $g(\widetilde{y}_{n+1}) = \text{sign}(k_1 - k_3)$, then:

$$\widehat{b}_2 = -\frac{1}{3} + 3\beta - \epsilon, \quad \widehat{b}_1 = \beta - \frac{\epsilon}{6}.$$

8. Else if $(k_2 = k_3)$ and sign $g(\widetilde{y}_{n+1}) = -\text{sign}(k_1 - k_3)$, then:

$$\widehat{b}_2 \text{ arbitrary}, \quad \widehat{b}_1 = \max\left\{\beta, \frac{1}{9} + \frac{\widehat{b}_2}{3}\right\} + \epsilon.$$

9. Else if $(k_1 + 3k_2 - 4k_3 = 0)$ and sign $g(\widetilde{y}_{n+1}) = -\text{sign}(k_2 - k_3)$, then:

$$\widehat{b}_1 \text{ arbitrary}, \quad \widehat{b}_2 = -\frac{1}{3} + 3\widehat{b}_1 - \frac{g(\widetilde{y}_{n+1})}{2h(k_2 - k_3)}.$$

Once obtained $\widehat{y}_{n+1}$, we will obtain the direction of the projection (4), and then, we will calculate the projected approximation y_{n+1} according to (3).

5 Numerical experiments

In this section, we are going to check the behaviour of the numerical method obtained in this paper projecting the 3rd-order Bogacki and Shampine RK method. We will apply it, together with the non projected method and the simplified standard projection, to some test problems. Thus:

- BS3 will denote the 3-stage 3rd-order RK method derived by Bogacki and Shampine, whose coefficients $(A, \widetilde{b}^T)$ are given in (8).
- pstBS3 will represent the projection method (3) obtained from the numerical solution $\widetilde{y}_{n+1}$ provided by BS3 by using simplified standard projection [16], which takes $w_n = \nabla g(\widetilde{y}_{n+1})$ as direction vector.

- pBS3 will denote the projection method (3) obtained from BS3 by following the results obtained in this paper. The direction of the projection w_n is taken according to (4), where the coefficients of the RK method $\widehat{y}_{n+1}$, embedded to $\widetilde{y}_{n+1}$, are calculated following the algorithm listed in the previous section. The free parameters in that algorithm have been taken equal to zero, and $\epsilon = 0.1$.

We have implemented these three methods with fixed step size. For the projection methods, the real parameter λ_n in (3) has been calculated by applying Newton iteration to the non linear equation

$$g(\widetilde{y}_{n+1} + \lambda_n w_n) = 0,$$

where $g(y) := G(y) - G(y_0)$, and $G(y)$ is a first integral of the differential system. All the figures in this paper are represented in a log-log scale.

First of all, we consider the harmonic oscillator, with frequency $\omega = 1$, to show numerically that, according to sections 3.1.1 and 3.1.2, the method obtained projecting BS3 has order 6 or order 2 depending on the order of the embedded method is $q = 1$ or $q = 2$, respectively. We have taken as coefficients of the embedded method of order $q = 1$ those given in item 3 of the algorithm described in section 4, and we have integrated the problem in the interval $[0, 624]$. For the embedded method of order $q = 2$, we have taken $\widehat{b}_1 = 0$ and $\widehat{b}_2 = 1$. Figure 1 shows clearly that the non projected method BS3 has order 3, whereas the projection pstBS3 has order 4, in agreement with the results in Corollary 2.1. On the other side, when the embedded method of order $q = 2$ is used in the projection, only order 2 is obtained, whereas the projection based on the first order embedded method attains order 6, as expected. Straight reference lines with slopes 6, 4, 3 and 2 have been drawn to make clear these results.

Next, we consider the Euler problem, which describes the evolution in time of the angular momentum $y = (y_1, y_2, y_3)^T$ of a free rigid body (see e.g. [16, pp. 95]):

$$\frac{d}{dt}\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 0 & c_3 y_3 & -c_2 y_2 \\ -c_3 y_3 & 0 & c_1 y_1 \\ c_2 y_2 & -c_1 y_1 & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}, \tag{19}$$

where $c_j^{-1} = I_j > 0$, $j = 1, 2, 3$, are the principal momenta of inertia. This is a Poisson system which has two first integrals

$$\begin{cases} E(y) = (c_1 y_1^2 + c_2 y_2^2 + c_3 y_3^2)/2 \quad \text{(Kinetic energy)}, \\ L(y)^2 = y_1^2 + y_2^2 + y_3^2 \quad \text{(Modulus of angular momentum)}. \end{cases} \tag{20}$$

By supposing that $c_1 > c_2 > c_3$, and given an initial value $y(0) = y_0$, the system (19) has a periodic solution with period T depending on the two

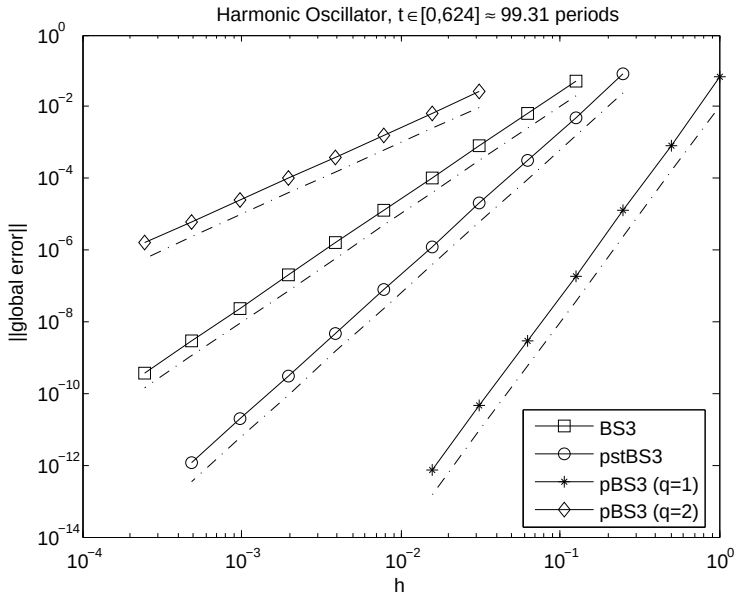

Figure 1: Harmonic oscillator, global error against step size, log-log scale

quadratic first integrals (20). More precisely, it is given by the elliptic integral

$$T = \frac{4}{\delta} K(k) = \frac{4}{\delta} \int_0^{\pi/2} \frac{dt}{\sqrt{1 - k^2 \sin^2 t}}, \tag{21}$$

with

$$k^2 = \frac{(c_1 - c_2)\big(2E(y_0) - c_3 L(y_0)^2\big)}{(c_2 - c_3)\big(c_1 L(y_0)^2 - 2E(y_0)\big)} < 1,$$

$$\delta = \big[(c_2 - c_3)\big(c_1 L(y_0)^2 - 2E(y_0)\big)\big]^{1/2}.$$

In our numerical experiments we haven taken $c_1 = 1/0.345$, $c_2 = 1/0.653$, $c_3 = 1$, which correspond to the water molecule, as considered in [13] and [25], together with the initial conditions $y(0) = (0.5, 0.2, \sqrt{1 - 0.5^2 - 0.2^2})$.

We have applied the projection techniques for this problem so that the function period given in (21) is preserved by the numerical solution. Figure 2 shows the global error obtained integrating over the interval $[0, 1208]$ against the step size for the Euler equations. Clearly, the two projection methods perform better than the basic formula BS3, being our new method pBS3 the most efficient of the three compared methods. It can be observed that, even though it has order three, it behaves as a 4th-order method, whereas the other two methods behave, as expected, as 3rd-order methods. Straight dash-dot lines with slopes 4 and 3 have been drawn there to show that behaviour.

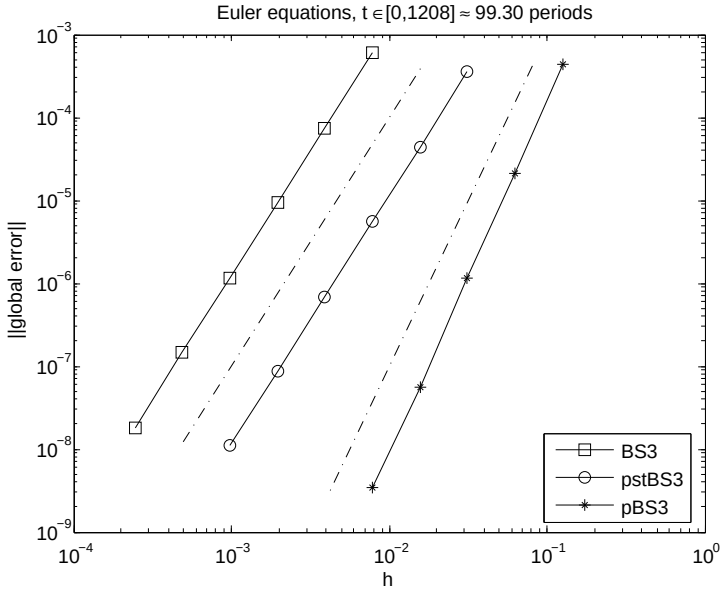

Figure 2: Euler equations, global error against step size, log-log scale

Our last test problem is the non forced Duffing equation [16]:

$$y''(t) + \omega^2 y(t) = k y(t)^3, \quad k > 0.$$

The energy function

$$H(y, y') = \omega^2 y^2 + y'^2 - k y^4/2,$$

is a first integral of this system, and we will only consider periodic solutions. For this case, it is known that the period depends only on the energy H. For a periodic motion, we solve for the smallest values in absolute value on the trajectory with $y' = 0$, i. e., $H(y, 0) = \text{cte.}$, obtaining $y^+ > 0$ and $y^- = -y^+$, and the expression for the period is given by

$$T = 2 \int_{y^-}^{y^+} \frac{1}{\sqrt{2(H - \omega^2 u^2/2 + k u^4/4)}} \, \mathrm{d}u.$$

In our numerical experiments we have taken the initial conditions $y(0) = 0$ and $y'(0) = \hat{\omega} = \sqrt{\omega^2 - k/2}$, with the parameters $w = 5$, $k = 0.1$ and an integration interval $[0, 125]$. With these values, the period is

$$T = 8\sqrt{\frac{5}{499}} \, \mathrm{elK}\left(\frac{1}{499}\right) = 1.258526506204981\ldots,$$

where elK(m) is the complete elliptic integral of the first kind. For this problem, projections have been done preserving this first integral H.

Figure 3 shows the superior efficiency of the projected methods with respect to the basic formula from which they come from. It can be observed in the figure that the projected methods perform as 4th-order methods for this problem. As in previous figures, discontinuous straight lines with slopes 3 and 4 have been drawn to make clear the order of the studied methods. The best behaviour corresponds again to our projected method pBS3 that provides a much smaller global error, due to its special properties for oscillatory problems.

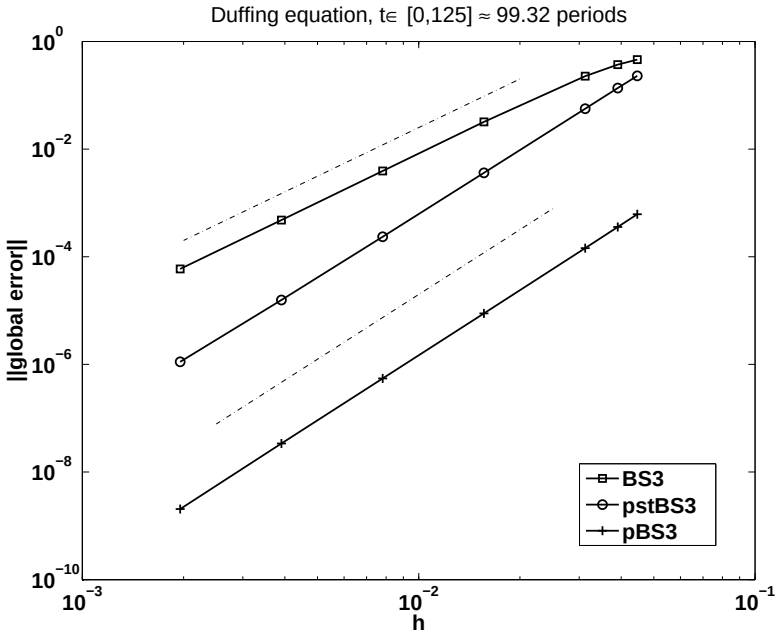

Figure 3: Duffing equation, global error against step size, log-log scale

6 Conclusions

In this paper we have analyzed the properties of the projections techniques in connection with Runge-Kutta methods when the direction of projection is obtained from suitable embedded Runge-Kutta schemes. We have obtained conditions on the projection so that the projected approximation attains higher order than the non projected solution. We have proved that by choosing properly the embedded formulae, the order can be highly increased for the harmonic oscillator. We have also given some criteria to select the embedded

scheme for general, non linear, problems and we have obtained particular projections for the third order method of Bogacki and Shampine. We have finally shown by means of several numerical experiments that the new proposed projection techniques are much more efficient than the standard, non projected schemes, and even than the orthogonal projection, mainly when problems with oscillatory behaviour are integrated.

References

[1] Bogacki, P., Shampine, L.F.: A 3(2) pair of Runge–Kutta formulas. Applied Mathematics Letters 2, no. 4, 321–325 (1989).

[2] Brugnano, L., Calvo, M., Montijano, J.I., Rández, L.: Energy-preserving methods for Poisson systems. J. Comput. Appl. Math. 236, no. 16, 3890–3904 (2012).

[3] Brugnano, L., Iavernaro, F.: Line integral methods which preserve all invariants of conservative problems. J. Comput. Appl. Math. 236, no. 16, 3905–3919 (2012).

[4] Brugnano, L., Iavernaro, F., Trigiante, D.: Hamiltonian BVMs (HBVMs): a family of "drift-free" methods for integrating polynomial Hamiltonian systems. AIP Conf. Proc. 1168, 715–718 (2009).

[5] Brugnano, L., Iavernaro, F., Trigiante, D.: A simple framework for the derivation and analysis of effective classes of one-step methods for ODEs. Appl. Math. Comput. 218, no. 17, 8475–8485 (2012).

[6] Calvo, M., Franco, J.M., Montijano, J.I., Rández, L.: Explicit Runge–Kutta methods for initial value problems with oscillating solutions. J. Comp. Appl. Math. 76, no. 1-2, 195–212 (1996).

[7] Calvo, M., Hernández–Abreu, D., Montijano, J.I., Rández, L.: On the preservation of invariants by explicit Runge–Kutta methods. SIAM J. Sci. Comput. 28, no. 3, 868–885 (2006).

[8] Calvo, M., Laburta, M.P., Montijano, J.I., Rández, L.: Approximate preservation of quadratic first integrals by explicit Runge–Kutta methods. Adv. Comput. Math. 32, no. 3, 255–274 (2010).

[9] Calvo, M., Laburta, M.P., Montijano, J.I., Rández, L.: Projection methods preserving Lyapunov functions. BIT Numer. Math. 50, no. 2, 223–241 (2010).

[10] Celledoni, E., McLachlan, R.I., McLaren, D.I., Owren, B., Quispel, G.R.W., Wright, W.M.: Energy preserving Runge–Kutta methods. M2AN Math. Model. Numer. Anal. 43, no. 4, 645–649 (2009).

[11] Cohen, D., Hairer, E.: Linear energy-preserving integrators for Poisson systems. BIT Numer. Math. 51, no. 1, 91–101 (2011).

[12] Cooper, G.J.: Stability of Runge–Kutta methods for trajectory problems. IMA J. Numer. Anal. 7, no. 1, 1–13 (1987).

[13] Fassò, F.: Comparison of splitting algorithm for the rigid body. J. Comput. Phys. 189, no. 2, 527–538 (2003).

[14] Grimm, V., Quispel, G.R.W.: Geometric integration methods that preserve Lyapunov functions. BIT 45, no. 4, 709–723 (2005).

[15] Hairer, E.: Energy-preserving variant of collocation methods. J. Numer. Anal. Ind. Appl. Math. (JNAIAM) 5, no. 1-2, 73–84 (2010).

[16] Hairer, E., Lubich, C., Wanner, G.: Geometric Numerical Integration: Structure Preserving Algorithms for Ordinary Differential Equations. Springer-Verlag, Berlin (2002).

[17] van der Houwen, P.J., Sommeijer, B.P.: Explicit Runge–Kutta (Nyström) methods with reduced phase errors for computing oscillating solutions. SIAM J. Numer. Anal. 24, no. 3, 595–617 (1987).

[18] van der Houwen, P.J., Sommeijer, B.P.: Phase-lag analyisis of implicit Runge–Kutta methods. SIAM J. Numer. Anal. 26, no. 1, 214–229 (1989).

[19] Iavernaro, F., Pace, B.: s-stage Trapezoidal Methods for the Conservation of Hamiltonian Functions of Polynomial Type. AIP Conf. Proc. 936, 603–606 (2007).

[20] Iavernaro, F., Trigiante, D.: High-order symmetric schemes for the energy conservation of polynomial Hamiltonian problems. J. Numer. Anal. Ind. Appl. Math. (JNAIAM) 4, no. 1-2, 87–101 (2009).

[21] McLachlan, R.I., Quispel, G.R.W.: Geometric integration of conservative polynomial ODEs. Appl. Numer. Math. 45, no. 4, 411–418 (2003).

[22] Quispel, G.R.W., Capel, H.W.: Solving ODE's numerically while preserving a first integral. Phys. Lett. A 218, no. 3-6, 223–228 (1996).

[23] Quispel, G.R.W., McLaren, D.I.: A new class of energy-preserving numerical integration methods. J. Phys. A: Math. Theor. 41, no. 4, 045206, 7 pp. (2008).

[24] Shampine, L.F., Reichelt, M.W.: The MATLAB ODE suite. SIAM J. Sci. Comput. 18, no. 1, 1–22 (1997).

[25] Vilmart, G.: Reducing round-off errors in rigid body dynamics. J. Comput. Phys. 227, no. 15, 7083–7088 (2008).