High order explicit Runge-Kutta pairs for ephemerides of the solar system and the Moon

Philip W. Sharp*

Department of Mathematics, University of Auckland, Private Bag 92019,

Auckland, NEW ZEALAND. sharp@math.auckland.ac.nz

December 9 1999

Abstract

Numerically integrated ephemerides of the solar system and the Moon require very accurate integrations of systems of second order ordinary differential equations. We present a new family of 8-9 pairs and assess the performance of two new 8-9 pairs on the equations used to create the ephemeris DE102. Part of this work is the introduction of these equations as a test problem for integrators of initial value ordinary differential equations.

1 Introduction

Ephemerides for the solar system and the Moon can be obtained by numerically integrating a system of ordinary differential equations which model all significant gravitational attractions between the bodies. To take full advantage of the accuracy of modern astronomical observations and to distinguish between competing analytical theories for the motion of the planets and the Moon, the global error in the integrations must be very small. Another characteristic of the integrations is that they often span a large interval of astronomical time, necessitating many integration steps.

The requirement of a very small global error and the need to take many steps often means the integrations must be performed using extended precision such as 80-bit arithmetic or quadruple precision.

The ordinary differential equations for ephemerides are non-stiff and hence explicit Runge-Kutta (ERK) pairs are suitable methods for performing the integrations. Of the many ERK pairs available, the 13-stage 7-8 pair of Prince and Dormand [5] has proven to be as efficient as any other on many problems when using double precision arithmetic, except possibly at low accuracy requirements.

*This work was supported by the University of Auckland Research Committee.

In particular, the pair is noticeably more efficient than 8-9 pairs. We investigate whether this result holds for numerically integrated ephemerides.

We begin our investigation with a search of an existing family of 16-stage 8-9 pairs for a near optimal pair. Next we derive a family of 17-stage 8-9 pairs and search this for a near optimal pair. We then compare the new pairs and the 7-8 pair of Prince and Dormand on the equations of DE102.

2 Order nine pairs

2.1 Definitions

Consider the initial value problem

$$y' = f(x, y), \quad y(x_0) = y_0, \tag{1}$$

where $' \equiv d/dx$, $f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ and the solution $y(x)$ is sufficiently differentiable.

The 8-9 ERK pairs we investigate have s-stages and generate an order nine approximation y_i and an order eight approximation $\widehat{y}_i$ to $y(x_i)$, $i = 1, 2, \ldots,$ according to

$$y_i = y_{i-1} + h \sum_{j=1}^{s} b_j f_j, \tag{2}$$

$$\widehat{y}_i = y_{i-1} + h \sum_{j=1}^{s} \widehat{b}_j f_j, \tag{3}$$

where $h = x_i - x_{i-1}$ and

$$f_j = f(x_{i-1} + hc_j, y_{i-1} + h \sum_{k=1}^{j-1} a_{jk} f_k), \quad j = 1, \ldots, s \quad (c_1 = 0).$$

We refer to $c_j, j = 1, \ldots, s$, as the abscissae and $a_{ij}, j = 1, \ldots, i-1, i = 2, \ldots, s$ as the interior weights. To retain the one step nature of the methods, we restrict the abscissae to the interval $[0, 1]$.

The efficiency of 8-9 pairs can be assessed using the size of the principal error coefficients. These can be written as

$$e_{10}(\tau_j) = \frac{\alpha(\tau_j)}{10!}(\gamma(\tau_j) \sum_{k=1}^{s} b_k \phi_k(\tau_j) - 1), \quad \tau_j \in T_{10},$$

where T_{10} is the set of rooted trees of order ten, $\alpha(t_j)$ and $\gamma(t_j)$ are positive integers and $\phi_k(\tau_j)$ are functions of the interior weights and abscissae. We use two measures of the size of the principal error coefficients

$$E_{10}^{2} = (\sum_{\tau_j \in T_{10}} e_{10}(\tau_j)^2)^{1/2}, \quad E_{10}^{\infty} = \max_{\tau_j \in T_{10}} \{|e_{10}(\tau_j)|\}. \tag{4}$$

2.2 Sixteen stages

Verner [7] derived a family of 16-stage 8-9 pairs with c_2, c_5, c_9, c_{10}, c_{11}, c_{13}, c_{14} and $a_{11,6}$ as free parameters (To simplify what follows, we have interchanged the coefficients for the fourteen and sixteenth stages, this can be done without changing the properties of the pairs.) The order nine formula in the pairs uses the first fifteen stages and the order eight formula uses all sixteen stages. The coefficients $b_j, \widehat{b}_j$, $j = 2, \ldots, 7$, b_{16}, $\widehat{b}_{14}$ and $\widehat{b}_{15}$ are identically zero.

Verner presented the coefficients of a pair from this family which had $c_2 = 1/12$, $c_5 = (2 + 2\sqrt{6})/15$, $c_9 = 1/2$, $c_{10} = 1/3$, $c_{11} = 1/4$, $c_{12} = 4/3$, $c_{13} = 5/6$, $c_{14} = 1/6$ and $a_{11,6} = 0$. The pair has $E^2_{10} = 6.1 \times 10^{-5}$ and $E^\infty_{10} = 3.1 \times 10^{-5}$ and has been used when comparing the numerical performance of 8-9 pairs with pairs of other orders. However the pair was intended as an illustration of the derivation and not as one with a near optimal value of E^2_{10} or E^∞_{10}.

To assess in a problem independent way if the 8-9 family of Verner contains more efficient pairs, and if so, how much more efficient, we performed a minimisation of E^2_{10} over the free parameters, subject to the constraint that the coefficients be no larger than M in magnitude. This constraint is commonly used when selecting a pair from a family and is intended to prevent the selection of a pair with poor round-off error properties. Although no one value of M is used, it is often 20 or 30 and we chose 30.

We performed the minimisation using an interactive grid search and obtained a minimum of 7.5×10^{-7} when working with a grid size of approximately 0.001. The pair we obtained had $c_2 = 0.020$, $c_5 = 0.311$, $c_9 = 0.312$, $c_{10} = 0.105$, $c_{11} = 0.105$, $c_{13} = 0.879$, $c_{14} = 0.916$ and $a_{11,10} = -0.150$ (as a matter of preference we have used $a_{11,10}$ in place of $a_{11,6}$ as a free parameter). The algorithms in [7] can be used to find the remaining coefficients.

A slightly smaller value of E^2_{10} is possible if a smaller grid size is used, but since the number of derivative evaluations varies approximately as the ninth root of E^2_{10}, the gain in efficiency is small. A significantly smaller value of E^2_{10}, approximately twice as small, is possible if the abscissae are not constrained to the interval $[0, 1]$, but this choice means the pair is no longer a one step method.

A comparison of E^2_{10} for the new pair and the one of Verner suggests the new pair will be approximately 70 percent more efficient at small stepsizes, raising the possibility of it being competitive with pairs of other orders.

2.3 Seventeen stages

The work of Sharp and Smart [6] for 4-5 and 5-6 ERK pairs shows a gain in efficiency is possible if an extra stage is used to form the pair. The extra stage means more free parameters are available, permitting a smaller value of E^2_{10}, but this is at the expense of increasing (by one) the number of function evaluations required to take a step.

To investigate if a gain in efficiency was possible for 8-9 pairs, we derived a family of 17-stage 8-9 pairs. The family has six more free parameters (three abscissae, three interior weights) than the 16-stage 8-9 pairs.

The key to the derivation is the vector of positive integers $\xi = [\xi_1, \xi_2, \ldots, \xi_{s-1}]^T$:

$$\frac{c_i^{k+1}}{k+1} = \sum_{j=1}^{i-1} a_{ij} c_j^k, \quad k = 0, \ldots, \xi_i - 1, \quad i = 1, \ldots, s-1 \tag{5}$$

$$a_{ij} = 0, \quad \text{if} \quad \xi_i > \xi_j + 1, \quad j = 1, \ldots, i-1, \quad i = 1, \ldots, s-1. \tag{6}$$

The 16-stage pairs have $\xi = [5, 1, 2, 3, 3, 4, 4, 5, 5, 5, 5, 5, 5, 5, 5]^T$; to obtain ξ for the 17-stage pairs, one positive integer less than five must be inserted. We examined three choices and found that inserting a 4 after the second 4 to give

$$\xi = [5, 1, 2, 3, 3, 4, 4, 4, 5, 5, 5, 5, 5, 5, 5, 5]^T$$

led to the largest number of free parameters.

With ξ specified, the derivation is similar to that for the 16-stage pairs, the main difference being fewer constraints on the abscissae for the first nine stages. We took c_2, c_5, c_6, c_7, c_9, c_{10}, c_{11}, c_{12}, c_{14}, c_{15}, $a_{8,7}$, $a_{11,10}$, $a_{12,10}$, $a_{12,11}$ as free parameters; other choices are possible, but the number of free parameters remains the same. The abscissae c_3, c_4, c_8, c_{16} and c_{17} are constrained as

$$c_3 = \frac{2}{3} c_4, \quad c_4 = \frac{3c_6 - 4c_5}{4c_6 - 6c_5} c_6, \quad c_8 = c_9 \frac{20c_6c_7 - 15c_6c_9 - 15c_7c_9 + 12c_9^2}{5(3c_9^2 - 4c_6c_9 + 6c_6c_7 - 4c_7c_9)}, \quad c_{16} = c_{17} = 1, \tag{7}$$

and the expression for c_{13} is the same as for c_{12} in the 16-stage pairs except c_8, c_9, c_{10} and c_{11} are replaced by c_9, c_{10}, c_{11} and c_{12} respectively.

We performed a minimisation of E_{10}^2 for the new family using an interactive grid search and steepest descent (a grid search by itself was impracticable because of the large number of free parameters) and obtained a pair with $E_{10}^2 = 1.0 \times 10^{-7}$ and $E_{10}^\infty = 3.6 \times 10^{-8}$. The value of the free parameters to four decimal places are $c_2 = 0.0757$, $c_5 = 0.3617$, $c_6 = 0.4139$, $c_7 = 0.1074$, $c_9 = 0.7607$, $c_{10} = 0.6068$, $c_{11} = 0.1531$, $c_{12} = 0.8333$, $c_{14} = 0.9733$, $c_{15} = 0.9888$, $a_{8,7} = -0.0001$, $a_{11,10} = -0.0078$, $a_{12,10} = 0.0067$ and $a_{12,11} = -0.0026$. Equations (7) together with ξ given above and the algorithms in [7] can be used to find the remaining coefficients.

A comparison of E_{10}^2 for the new 16-stage and 17-stage pairs suggests, after scaling by the number of stages, the 17-stage pair will be 18 percent more efficient at small stepsizes.

2.4 Generalised

The families of 8-9 pairs described in the previous sub-section are readily generalised to include either one or two extra free parameters.

One generalisation is to replace $\widehat{b}_j$, $j = 1, \ldots, s$ by the convex linear combination $\alpha b_j + (1-\alpha)\widehat{b}_j$. This substitution is equivalent to making one of the previously identically zero $\widehat{b}$ a free parameter. The local error estimate for the pair is changed, but since b_j,

$j = 1, \ldots, s-1$ remain the same, the error coefficients of the order nine formulae and hence E^2_{10} (and E^{∞}_{10}) are unchanged.

The second generalisation is based on a transformation obtained by Verner [8] for two families of 8-stage 5-6 (ERK) pairs. Verner showed the family of Prince and Dormand [5] which has c_2, c_3, c_5, c_6, b_8 and $\widehat{b}_7$ as free parameters can be obtained from the class in Verner [7] which has c_2, c_3, c_5 and c_6 as free parameters using a simple transformation on the last two rows of interior weights.

This transformation generalises (Verner, private communication) to other families of pairs, including the 8-9 pairs in this paper. This means b_{16} and $\widehat{b}_{15}$, previously zero in the 16-stage 8-9 pairs, and b_{17} and $\widehat{b}_{16}$, previously zero in the 17-stage 8-9 pairs, can be free parameters.

The introduction of these two free parameters changes the local error estimate and the principal error coefficients of the order nine formula. However, as is the case for the 5-6 pairs in [7], the change in E^2_{10} and E^{∞}_{10} is small for near optimal pairs.

3 DE102

Newhall, Standish and Williams [4] presented DE 102, an ephemeris of the solar system and the Moon, obtained by integrating a system of 33 second order ordinary differential equations of the form

$$y'' = f(t, y, y'). \tag{8}$$

The system (8) consists of equations of motion for the nine planets, the Moon together with three equations for the lunar physical librations. The motion of the Sun is found from the definition of the solar system barycentre. The equations contain contributions from point-mass interactions, figure effects for Earth and the Moon, Earth tides and perturbations from the five asteroids (1) Ceres, (2) Pallas, (4) Vesta, (7) Iris and (324) Bamberga.

The calculations required for one evaluation of the second derivative for (8) are described in Figure 1. A fuller description is given in [4] and by inference in the program DE118i.ARC of Moshier, available on the internet.

4 Numerical experiments

We conducted numerical tests of the two new 8-9 pairs and the 7-8 pair of Prince and Dormand on the DE102 equations described in the previous section. The results are illustrated below. The pairs are denoted by PD78 (Prince and Dormand 7-8), P16 (new 16-stage) and P17 (new 17-stage).

A computer which performed quadruple precision in hardware was unavailable and hence we used the the multiprecision Fortran90 package MPFUN90 of Bailey [1], with the precision level set at 35 digits, approximately that of quadruple precision. The multi-

1. Initialise

a) Calculate the heliocentric position and velocity for the asteroids and transform to approximate barycentric values. These values are corrected once the correct position of the Sun is known.
b) Calculate the distance between the bodies. The distances involving the Sun or asteroids are estimates only. These distances are corrected once the correct position of the Sun is known.
c) Use fixed-point iteration to find the correct position and velocity of the Sun and asteroids, then correct the distances involving the Sun or asteroids.
d) Calculate the cube of the distances between all bodies.

2. Point-mass acceleration

a) Calculate the Newtonian acceleration of all bodies.
b) Calculate the post-Newtonian acceleration of the planets and the Moon.

3. Figure of the Moon

a) Form the rotation matrix for the transformation from space to body coordinates.
b) Calculate the effect of the point-mass Earth on the lunar figure and add this to the lunar acceleration.
c) Calculate the torque on the Moon due to the gravitational interaction between the lunar figure and the external point-mass Earth.
d) The acceleration from b) induces an acceleration in the Earth - add this to the Earth's acceleration.
e) Calculate the effect of the point-mass Sun on the lunar figure and add this to the lunar acceleration.
f) Calculate the torque on the Moon due to the gravitational interaction between the lunar figure and the external point-mass Sun.
g) Calculate the acceleration of the libration angles.

4. Figure of the Earth

a) Calculate the effect of the point-mass Moon on the Earth's figure and add this to the Earth's acceleration.
b) The acceleration from a) induces an acceleration in the Moon - add this to the lunar acceleration.
c) Calculate the contribution to the acceleration of the Moon and the Earth due to the Earth tides.
d) Calculate the effect of the point-mass Sun on the Earth's figure and add this to the Earth's acceleration.
The accelerations in this section are adjusted for the precession and nutation of the equinox and obliquity of the ecliptic.

Figure 1: A summary of the calculations required for one evaluation of the second derivative in the mathematical model of DE 102.

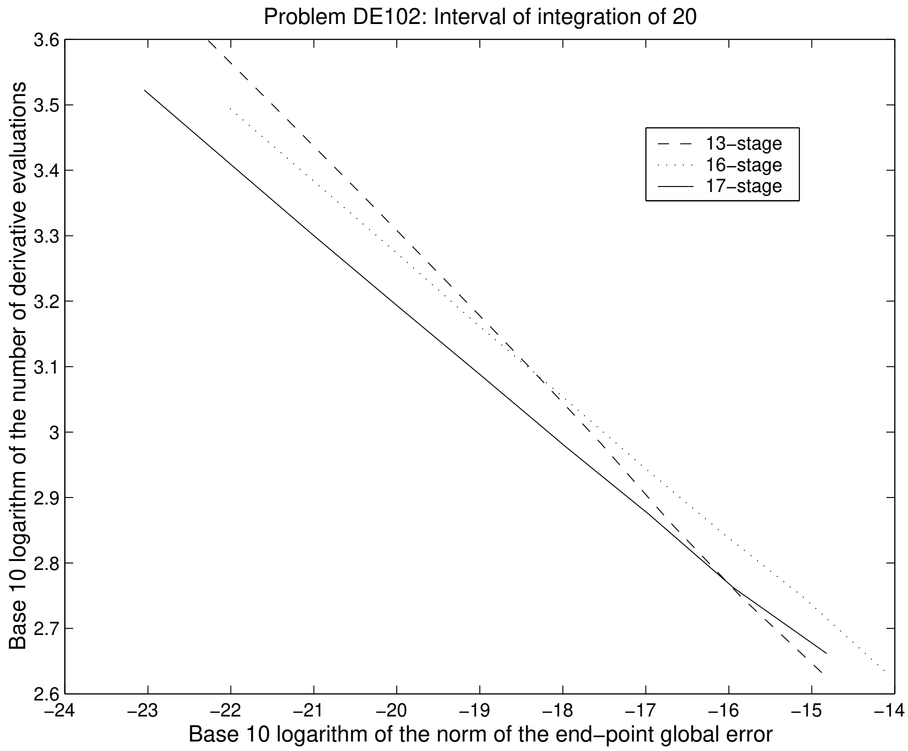

Figure 2: A log-log graph of the number of derivative evaluations against the norm of the end-point global error for DE102 with a integration interval of 20. Prince and Dormand 7-8 pair - dashed line, new 16-stage pair - dotted line, new 17-stage pair - solid line.

precision version of our program was 270 times slower than our double precision version which makes the use of MPFUN90 impractical for long integrations.

The coefficients of the 7-8 pair as specified in [5] are accurate to approximately 18 digits. We recalculated the coefficients in 100 digit arithmetic, using the values of the free parameters in [5], and used these coefficients, rounded to 35 digits. The global error in a numerical solution was obtained by calculating a more accurate solution and taking the difference between the two solutions.

The first example is for an integration interval of 20 and local error tolerances of 10^i, $i = -14, \ldots, -22$. Figure 2 contains the log-log graph of the number of derivative evaluations against the norm of the end-point global error (the points have been joined for clarity). Pair P17 is more efficient than P16 suggesting the efficiency is improved by adding a stage. The gain in efficiency varies from 15 to 20 percent, in good agreement with that predicted using E_{10}^2. The pairs P16 and P17 are more efficient than PD8 for global errors smaller than (approximately) 10^{-16}, and $10^{-18.5}$ respectively. In addition and as can be expected from the order of the pairs, the efficiency of 8-9 pairs relative to the 7-8 pair increases as the global error decreases. For example, P17 is 16 percent more efficient for a global error of 10^{-20} and 29 percent more efficient for a global error of 10^{-22}.

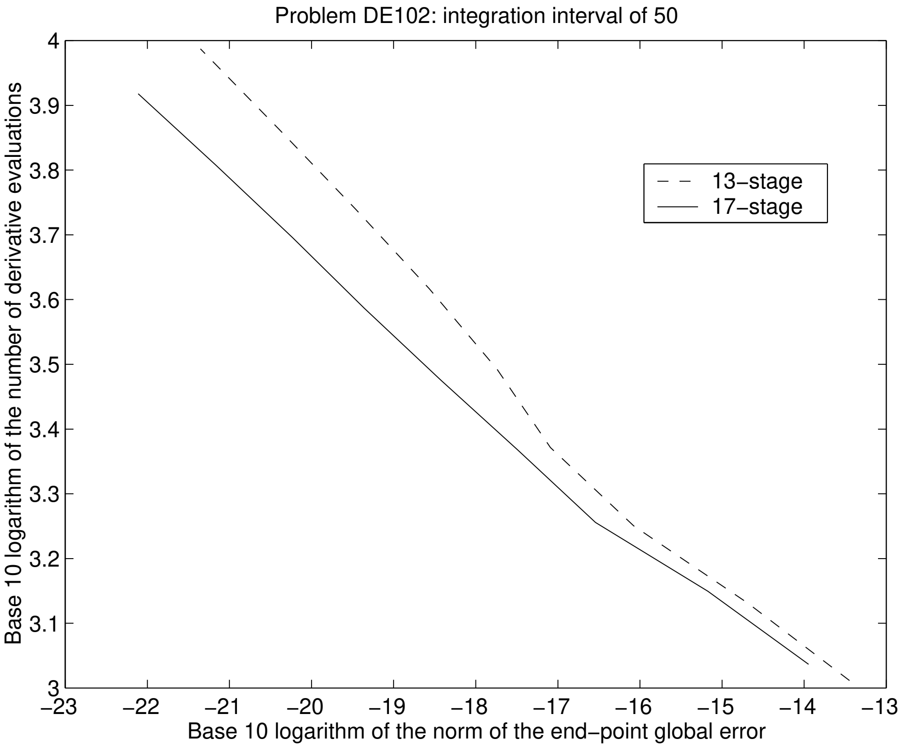

Figure 3: A log-log graph of the number of derivative evaluations against the norm of the end-point global error for DE102 with a integration interval of 50. Prince and Dormand 7-8 pair - dashed line, new 17-stage pair - solid line.

The second example is for an integration interval of 50 using the same local error tolerances as in the first example. The results are given in Figure 3. P16 was excluded because our test results such as those in Figure 2 showed P16 was less efficient than P17 for the local error tolerances we were using.

The efficiency of P17 relative to PD78 as a function of the global error is similar to that for the first example, except for a minor difference at the larger global errors. The global errors are larger than in the first example, a result which is consistent with a larger interval of integration.

5 Discussion

The main aim of our work was to investigate if 8-9 explicit Runge-Kutta pairs were more efficient than lower order pairs, principally 7-8 pairs, for numerically integrated ephemerides. We derived a new family of 8-9 pairs, obtained near optimal 8-9 pairs from this family and an existing one, and compared the performance of these pairs and the 7-8 pair of Prince and Dormand on the equations of the ephemerides DE102.

Our testing showed the 8-9 pairs were usually more efficient than the 7-8 pairs. The gain in efficiency was not large, but given the amount of CPU time required to produce ephemerides, the gain is significant. Our testing also showed that near optimal 17-stage 8-9 pairs can be more efficient than near optimal 16-stage 8-9 pairs.

As part of this work we introduced the equations of DE102 as a test problem for integrators of initial value ordinary differential equations. This problem, in addition to being a realistic one, has several interesting numerical aspects. For example, the position and velocity of the Sun is found by solving a system of nonlinear (algebraic) equations. The question arises as to the most efficient way of solving these equations.

References

[1] D.H. Bailey, *A Fortran-90 based multiprecision System*, RNR Technical Report RNR-94-013, NAS Scientific Computation Branch, NASA Ames Research Center, January, 1995.

[2] E. Fehlberg, *Classical fifth-,sixth-,seventh-,and eighth-order Runge-Kutta formulas with stepsize control*, NASA Technical Report NASA TR R-287 (1968), 82 pages.

[3] S.L. Moshier, *Comparison of a 7000-year lunar ephemeris with analytical theory*, Astron. Astrophys. **262** (1982), 613-616.

[4] X.X. Newhall, E.M. Standish, J.G. Williams, *DE 102: a numerically integrated ephemeris of the Moon and planets spanning forty-four centuries*, Astron. Astrophys. **125** (1983), 150-167.

[5] P.J. Prince and J.R. Dormand, *High-order embedded Runge-Kutta formulae*, J. Comput. Appl. Math. **7** (1981), 67-76.

[6] P.W. Sharp, E. Smart, *Explicit Runge-Kutta pairs with one more derivative evaluation than the minimum*, SIAM J. Sci. Comput. **14** (1993), 338-348.

[7] J.H. Verner, *Explicit Runge-Kutta methods with estimates of the local truncation error*, SIAM J. Num. Anal. **15** (1978), 772-790.

[8] J.H. Verner, *A contrast of some Runge-Kutta formula pairs*, SIAM J. Num. Anal. **27** (1990), 1332-1344.