Càlcul numèric (2020-2021)

Numerical integration of ODEs

1 Introduction

1.1 The initial value problem. Euler's method.

The initial value problem (IVP), or Cauchy's problem, is to find, in an interval $[t_0, t_f]$, the solution of a system of ODEs verifying an initial condition,

$$\left.\begin{array}{r} y'(t) = f(t, y(t)), \\ y(t_0) = y_0 \end{array}\right\}, \quad t \in [t_0, t_f],$$

with $y(t), f(t, y(t)) \in \mathbb{R}^m$.

The existence and uniqueness theorem for the IVP says that if f is continuous with respect to t and there is a Lipschitz constant, L, in $[t_0, t_f] \times \Omega$, where Ω is a domain in $\mathbb{R}^m$, such that

$$\|f(t, y) - f(t, z)\| \leq L \|y - z\|, \qquad \text{in} \quad [t_0, t_f] \times \Omega,$$

then there is a unique solution of the IVP.

Consider now the integral form in the interval $[t_n, t_{n+1}]$

$$y(t_{n+1}) = y(t_n) + \int_{t_n}^{t_{n+1}} y'(t)\, dt = y(t_n) + \int_{t_n}^{t_{n+1}} f(t, y(t))\, dt \tag{1}$$

and approximate the integral by an adequate quadrature formula. For instance, Euler's method

$$y_{n+1} = y_n + h_n f(t_n, y_n), \qquad n = 0, 1, \ldots, N-1,$$

where $h_n = t_{n+1} - t_n, \qquad n = 0, 1, \ldots, N-1$, $y_0 = y(t_0)$ and y_n means the approximation to $y(t_n)$, is obtained by the rectangle formula

$$\int_{t_n}^{t_{n+1}} f(t, y(t))\, dt \approx \int_{t_n}^{t_{n+1}} f(t_n, y(t_n))\, dt = h_n f(t_n, y(t_n)),$$

in which the integrand is approximated by the value at the lower end of the interval of integration

$$f(t, y(t)) \approx f(t_n, y(t_n)) \approx f(t_n, y_n), \qquad t \in [t_n, t_{n+1}].$$

Euler's method can be improved by approximating in (1)

$$f(t, y(t)) \approx f(t_{n+1/2}, y(t_{n+1/2})), \qquad t \in [t_n, t_{n+1}],$$

where $t_{n+1/2} = t_n + h_n/2 = (t_n + t_{n+1})/2$. But we don't know $y(t)$ at $t_{n+1/2}$. We can approximate it by Euler's method with a time step $h_n/2$, i.e.,

$$f(t_n + \frac{h_n}{2}, y(t_n + \frac{h_n}{2})) \approx f(t_n + \frac{h_n}{2}, y_n + \frac{h_n}{2} f(t_n, y_n)).$$

This is the *improved Euler's method* due to Runge

$$\begin{aligned} y_{n+1} &= y_n + h_n k_2, \\ \\ k_1 &= f(t_n, y_n), \\ k_2 &= f(t_n + \tfrac{h_n}{2}, y_n + \tfrac{h_n}{2} k_1). \end{aligned}$$

1.2 Explicit Runge-Kutta methods.

The Runge-Kutta methods were first described by Carl Runge (1895) and Martin Kutta (1905). The idea of the improved Euler's method of using quadrature formulas in $[t_n, t_{n+1}]$, and obtaining the values of the integrand at the nodes by Euler's method, is the base of other Runge-Kutta methods. For instance, the *classical* fourth order Runge-Kutta method has the form

$$\begin{aligned} y_{n+1} &= y_n + \tfrac{h_n}{6}(k_1 + 2k_2 + 2k_3 + k_4), \\ \\ k_1 &= f(t_n, y_n), \\ k_2 &= f(t_n + \tfrac{h_n}{2}, y_n + \tfrac{h_n}{2} k_1), \\ k_3 &= f(t_n + \tfrac{h_n}{2}, y_n + \tfrac{h_n}{2} k_2), \\ k_4 &= f(t_n + h_n, y_n + h_n k_3). \end{aligned}$$

To approximate the integral in (1), it uses Simpson's rule,

$$\int_{t_n}^{t_{n+1}} f(t, y(t))\, dt \approx \frac{h_n}{6}\Big(f(t_n, y(t_n)) + 4f(t_n + \frac{h_n}{2}, y(t_n + \frac{h_n}{2})) + f(t_{n+1}, y(t_{n+1}))\Big),$$

and approximates the values of the integrand as

$$f(t_n + \frac{h_n}{2}, y(t_n + \frac{h_n}{2})) \approx \frac{1}{2}(k_2 + k_3), \qquad f(t_{n+1}, y(t_{n+1})) \approx k_4.$$

Other Runge-Kutta methods are, for instance, *Heun's method*,

$$\begin{aligned} y_{n+1} &= y_n + \tfrac{h_n}{4}(k_1 + 3k_3), \\ \\ k_1 &= f(t_n, y_n), \\ k_2 &= f(t_n + \tfrac{h_n}{3}, y_n + \tfrac{h_n}{3} k_1), \\ k_3 &= f(t_n + \tfrac{2}{3} h_n, y_n + \tfrac{2}{3} h_n k_2), \end{aligned}$$

or the *3/8 rule* ,

$$\begin{aligned} y_{n+1} &= y_n + \tfrac{h_n}{8}(k_1 + 3k_2 + 3k_3 + k_4), \\ \\ k_1 &= f(t_n, y_n), \\ k_2 &= f(t_n + \tfrac{h_n}{3}, y_n + \tfrac{h_n}{3}k_1), \\ k_3 &= f(t_n + \tfrac{2}{3}h_n, y_n + \tfrac{h_n}{3}(3k_2 - k_1)), \\ k_4 &= f(t_n + h_n, y_n + h_n(k_1 - k_2 + k3)). \end{aligned}$$

All these are examples of *explicit Runge-Kutta methods of s stages*, which have the general form

$$\begin{aligned} y_{n+1} &= y_n + h_n(b_1k_1 + \cdots + b_sk_s), \\ \\ k_1 &= f(t_n, y_n), \\ k_2 &= f(t_n + c_2h_n, \quad y_n + h_na_{21}k_1), \\ k_3 &= f(t_n + c_3h_n, \quad y_n + h_n(a_{31}k_1 + a_{32}k_2)), \\ &\vdots \qquad\qquad \vdots \\ k_s &= f(t_n + c_sh_n, \quad y_n + h_n(a_{s1}k_1 + \cdots + a_{s,s-1}k_{s-1})). \end{aligned} \tag{2}$$

The Euler's, improved Euler's, Heun's, the classical Runge-Kutta, and the 3/8 rule methods have, respectively, 1,2,3,4, and 4 stages.

In 1964, the New Zealander J. C. Butcher introduced what, in his honor, is known as Butcher's tableau,

$$\begin{array}{c|ccccc} 0 & 0 & & & & \\ c_2 & a_{21} & & & & \\ c_3 & a_{31} & a_{32} & & & \\ \vdots & \vdots & \vdots & \ddots & & \\ c_s & a_{s1} & a_{s2} & \ldots & a_{s,s-1} & \\ \hline & b_1 & b_2 & \ldots & b_{s-1} & b_s \end{array}$$

For instance, the tableau of Euler, improved Euler, and classical Runge-Kutta are

Euler

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}$$

Improved Euler

$$\begin{array}{c|cc} 0 & 0 & \\ \frac{1}{2} & \frac{1}{2} & \\ \hline & 0 & 1 \end{array}$$

Classical RK

$$\begin{array}{c|cccc} 0 & 0 & & & \\ \frac{1}{2} & \frac{1}{2} & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & & \\ 1 & 0 & 0 & 1 & \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array}$$

Except for some examples of little practical importance, all Runge-Kutta methods satisfy

$$\sum_j a_{ij} = c_i, \qquad i = 1, 2, \ldots, s.$$

We will assume this is the case in the following.

All explicit Runge-Kutta methods can be written as

$$y_{n+1} = y_n + h_n \Phi(t_n, y_n, h_n), \qquad n = 0, 1, \ldots, N-1,$$

where Φ depends of f. The same expression describes many other integration methods.

1.3 Global and local errors. Convergence

Given an IVP

$$\left.\begin{array}{c} y'(t) = f(t, y(t)), \\ y(t_0) = y_0 \end{array}\right\}, \quad t \in [t_0, t_f]. \tag{3}$$

and a numerical method of the form

$$y_{n+1} = y_n + h_n \Phi(t_n, y_n, h_n), \qquad n = 0, 1, \ldots, N-1, \tag{4}$$

on a partition

$$t_0 < t_1 < \ldots < t_N = t_f,$$

we define the *global error* or, simply, the error, to the differences

$$y(t_n) - y_n, \qquad n = 0, 1, \ldots, N,$$

i.e., to the difference between the numerical approximations, and the values of the exact solution of (3) at the corresponding points in the partition of $[t_0, t_f]$.

It is said that the method (4) is *convergent* if for any IVP (3) with regular enough f

$$\max_{0 \le n \le N} \|y(t_n) - y_n\| \to 0, \qquad \text{when} \quad h = \max_{0 \le n \le N-1} h_n \to 0,$$

and it is said to be *convergent of order p* when, in addition,

$$\max_{0 \le n \le N} \|y(t_n) - y_n\| = O(h^p).$$

The value of p is called the *order of convergence* of the method.

Remark 1 It is important to note that, as $t_N = t_f$, when $h \to 0$, the $N \to \infty$. For instance, in the case of a uniform mesh, $t_f = t_0 + Nh$, and $N = (t_f - t_0)/h$. Then in most cases one can write $\max_{t_0 \le t_n \le t_f} \|y(t_n) - y_n\|$, which is often more clear.

Euler's method has order $p = 1$. The orders of convergence of the other Runge-Kutta methods we have seen are

- **Improved Euler's method**, $p = 2$.
- **Heun's method**, $p = 3$.
- **Classical Runge-Kutta method**, $p = 4$.
- **3/8 rule method**, $p = 4$.

Finding bounds for the global errors is not difficult for methods of the form $y_{n+1} = y_n + h_n\Phi(t_n, y_n, h_n)$. The reason for which the errors decrease with h can be understood if we look at Fig. 1. The values y_n only agree, in general, with $y(t_n)$ at the first point $n = 0$. For the rest of the points y_n, a different solution passes with different initial condition at t_0 than y_0. In a step from y_n to y_{n+1}, this deviates from the solution $z_n(t)$ of the IVP

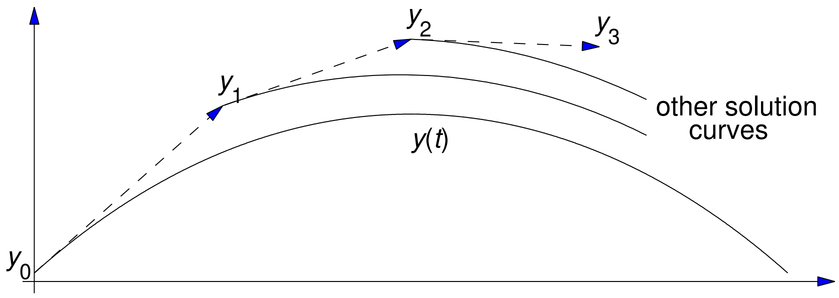


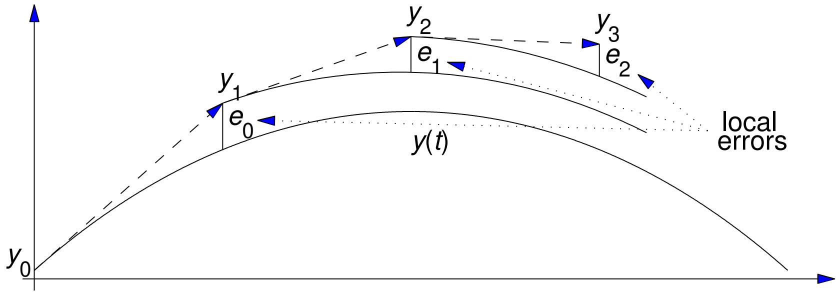
Figure 1:

$$\left.\begin{array}{c} z'(t) = f(t, z(t)), \\ z(t_n) = y_n \end{array}\right\}, \quad t \in [t_0, t_f]. \tag{5}$$

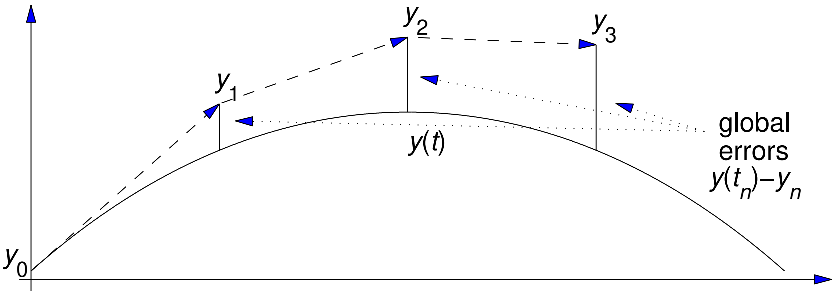
The *local error* at t_n is the difference

$$e_n = z_n(t_{n+1}) - y_{n+1} = z_n(t_{n+1}) - y_n - h_n\Phi(t_n, y_n, h_n), \qquad n = 0, 1, \ldots, N-1,$$

i.e., the error at t_{n+1} relative to the solution $z(t)$ of 5 which satistied $z(t_n) = y_n$.



This must not be confused with the global error $y(t_n) - y_n$.



It can be seen that

$$\max_{0\le n\le N} \|y(t_n) - y_n\| \le \frac{e^{L(t_f-t_0)} - 1}{L} \max_{0\le n\le N-1} \left\|\frac{e_n}{h_n}\right\|, \qquad (6)$$

L being the Lipschitz constant of f. Then we see that *a numerical method of the type* $y_{n+1} = y_n + h_n\Phi(t_n, y_n, h_n)$ *is convergent of order* p *if its local errors are* $O(h^{p+1})$.

When the local errors are $O(h^{p+1})$, it is said that the method is consistent of order p. For the Runge-Kutta methods, consistency and convergence are equivalent concepts. This is no so, for instance, for multistep methods.

The local error at t_n is similar to that at t_0 but we must change the solution z_n of (5) which satisfies $z_n(t_n) = y_n$ by the solution y of (3) which we want to approximate. Then it is enough to study the local error at t_0 to know the order of convergence of a method.

Example 1 Let study the local error of Euler's method, $y_1 = y_0 + h_0 f(t_0, y_0)$,

$$e_0 = y(t_1) - y_1 = y(t_1) - (y_0 + h_0 f(t_0, y_0)) = y(t_0 + h) - (y(t_0) + hy'(t_0)). \qquad (7)$$

The Taylor expansion of $y(t)$ at t_0 is

$$y(t_0 + s) = y(t_0) + sy'(t_0) + \frac{s^2}{2}y''(\xi),$$

for some $\xi \in (t_0, t_0 + s)$. Then

$$e_0 = \frac{h_0^2}{2}y''(\xi), \qquad \Rightarrow \qquad \|e_0\| = O(h_0^2) = O(h^2).$$

Note 1 The local errors of the Runge-Kutta methods have not, in general, an easy expression as in Euler's method case. We will denote them as $\Psi(y_n)$, i.e.,

$$e_n = h_n^{p+1}\Psi(y_n) + O(h^{p+2}),$$

understanding that each method has its own function Ψ.

1.4 Embedded Runge Kutta methods. Time step control

It is clear that it is not practical to take all the steps h_n of the same size. There might be points at which the function $\Psi(y_n)$ takes much higher values than at others. At the former h_n must be smaller that at the latter.

The most efficient way of controlling the time step, in the case of Runge-Kutta methods, is based on the *embedded Runge-Kutta pairs.* These pairs are formed by two Runge-Kutta methods which share the coefficients c and A.

$$\begin{array}{c|c} c & A \\ \hline & b^T \\ \hline & \hat{b}^T \end{array} \qquad \text{with} \quad \begin{array}{c|c} c & A \\ \hline & b^T \end{array} \quad \text{of order } p, \qquad \text{and} \quad \begin{array}{c|c} c & A \\ \hline & \hat{b}^T \end{array} \quad \text{of order } \hat{p}.$$

Examples of embedded pairs are the Runge-Kutta-Fehlberg 2(3), RKF23B, due to Fehlberg in 1968, with tableau

$$\begin{array}{c|cccc}
0 & 0 & & & \\
\frac{1}{4} & \frac{1}{4} & & & \\
\frac{27}{40} & -\frac{189}{800} & \frac{729}{800} & & \\
1 & \frac{214}{891} & \frac{1}{33} & \frac{650}{891} & \\
\hline
b_i & \frac{214}{891} & \frac{1}{33} & \frac{650}{891} & 0 \\
\hline
\hat{b}_i & \frac{533}{2106} & 0 & \frac{800}{1053} & -\frac{1}{78}
\end{array}$$

with orders $p = 2$ and $\hat{p} = 3$, the Runge-Kutta-Fehlberg 4(5), RKF45, due to Fehlberg in 1969, with tableau

$$\begin{array}{c|cccccc}
0 & 0 & & & & & \\
\frac{1}{4} & \frac{1}{4} & & & & & \\
\frac{3}{8} & \frac{3}{32} & \frac{9}{32} & & & & \\
\frac{12}{13} & \frac{1932}{2197} & -\frac{7200}{2197} & \frac{7296}{2197} & & & \\
1 & \frac{439}{216} & -8 & \frac{3680}{513} & -\frac{845}{4104} & & \\
\frac{1}{2} & -\frac{8}{27} & 2 & -\frac{3544}{2565} & \frac{1859}{4104} & -\frac{11}{40} & \\
\hline
b_i & \frac{25}{216} & 0 & \frac{1408}{2565} & \frac{2197}{4104} & -\frac{1}{5} & 0 \\
\hline
\hat{b}_i & \frac{16}{135} & 0 & \frac{6656}{12825} & \frac{28561}{56430} & -\frac{9}{50} & \frac{2}{55}
\end{array}$$

and with $p = 4$ and $\hat{p} = 5$, and DOPRI54, due to Dormand and Prince in 1980 with tableau

$$\begin{array}{c|ccccccc}
0 & 0 \\
\frac{1}{5} & \frac{1}{5} \\
\frac{3}{10} & \frac{3}{40} & \frac{9}{40} \\
\frac{4}{5} & \frac{44}{45} & -\frac{56}{15} & \frac{32}{9} \\
\frac{8}{9} & \frac{19372}{6561} & -\frac{25360}{2187} & \frac{64448}{6561} & -\frac{212}{729} \\
1 & \frac{9017}{3168} & -\frac{355}{33} & \frac{46732}{5247} & \frac{49}{176} & -\frac{5103}{18656} \\
1 & \frac{35}{384} & 0 & \frac{500}{1113} & \frac{125}{192} & -\frac{2187}{6784} & \frac{11}{84} \\
\hline
b_i & \frac{35}{384} & 0 & \frac{500}{1113} & \frac{125}{192} & -\frac{2187}{6784} & \frac{11}{84} & 0 \\
\hline
\hat{b}_i & \frac{5179}{57600} & 0 & \frac{7571}{16695} & \frac{393}{640} & -\frac{92097}{339200} & \frac{187}{2100} & \frac{1}{40}
\end{array}$$

and with $p = 5$ and $\hat{p} = 4$.

With an embedded pair, the method of order p is used to compute the new approximation, and the method of order $\hat{p}$ to estimate the local error as we will now show. Originally, as for the RKF23B pair, $\hat{p} > p$ was taken because the estimation of the error was more precise, but it is nowadays common to take $p > \hat{p}$ because it has been shown that this is, in practice, a more efficient option.

Obtaining $\hat{y}_{n+1} = y_n + h_n(\hat{b}_1 k_1 + \cdots + \hat{b}_s k_s)$ once y_{n+1} has been obtained is cheap, because it consists only in combining linearly the stages k_i already computed. If $\hat{p} < p$, we have, for the local errors,

$$\begin{aligned}
z(t_{n+1}) - y_{n+1} &= \Psi(y_n) h_n^{p+1} + O(h_n^{p+2}), \\
z(t_{n+1}) - \hat{y}_{n+1} &= \hat{\Psi}(y_n) h_n^{\hat{p}+1} + O(h_n^{\hat{p}+2}),
\end{aligned}$$

and by substracting,

$$\hat{y}_{n+1} - y_{n+1} = -\hat{\Psi}(y_n) h_n^{\hat{p}+1} + O(h_n^{\hat{p}+2}).$$

Then we have, as dominant term, the local error of the method of the lower order $\hat{p}$. The norm

$$\mathtt{EST}_n = \|\hat{y}_{n+1} - y_{n+1}\| \approx \|z(t_{n+1}) - \hat{y}_{n+1}\|,$$

is taken as an estimation of the local error at t_n, and it is compared with a tolerance TOL. It it is greater, the computed value y_{n+1} is rejected and a new step is taken with a different h_n. How is the new h_n chosen?

We have that with h_n

$$\mathtt{EST}_n \approx \|z(t_n + h_n) - \hat{y}_{n+1}(h_n)\| \approx \left\|\hat{\Psi}(y_n)\right\| h_n^{\hat{p}+1} > \mathtt{TOL},$$

and we want that with h'_n,

$$\mathtt{EST}'_n \approx \|z(t_n + h'_n) - \hat{y}_{n+1}(h'_n)\| \approx \left\|\hat{\Psi}(y_n)\right\| h_n'^{\hat{p}+1} = \mathtt{TOL},$$

then

$$h'_n = \sqrt[\hat{p}+1]{\frac{\mathtt{TOL}}{\|\hat{\Psi}(y_n)\|}} = h_n \sqrt[\hat{p}+1]{\frac{\mathtt{TOL}}{\|\hat{\Psi}(y_n)\|h_n^{\hat{p}+1}}} \approx h_n \sqrt[\hat{p}+1]{\frac{\mathtt{TOL}}{\mathtt{EST}_n}}.$$

Therefore the new step is taken with

$$h'_n = 0.9 h_n \sqrt[\hat{p}+1]{\frac{\mathtt{TOL}}{\mathtt{EST}_n}},$$

where the factor 0.9 is introduced for security reasons, because we have neglected the high order terms in the error.

If, on the contrary, $\mathtt{EST}_n \leq \mathtt{TOL}$, the value of y_{n+1} computed is accepted and the method proceeds to compute y_{n+2} after computing an h_{n+1} such that in the next step the local error is $\mathtt{TOL}$, i.e.,

$$h_{n+1} = \min(\, h_{\max}\, ,\, \mathtt{facmax}\, h_n\, ,\, 0.9 h_n \sqrt[\hat{p}+1]{\frac{\mathtt{TOL}}{\mathtt{EST}_n}} \quad),$$

where, also for security, the step h_{n+1} is limited by a maximal step h_{max}, and it is now allowed to grow above the previous step by a factor $\mathtt{facmax}$.

It is common to estimate the error by a combination of the absolute and relative errors, then

$$\mathtt{TOL} = \mathtt{ATOL} + \mathtt{RTOL}\, \|y_{n+1}\|\, , \quad \text{or} \quad \mathtt{TOL} = \max\left(\mathtt{ATOL}, \mathtt{RTOL}\, \|y_{n+1}\|\right)$$

or similar expressions but for each component of the vector $y_{n+1} \in \mathbb{R}^m$, for instance

$$\mathtt{TOL} = \max_{i=1,\cdots,m} \left(\mathtt{ATOL(i)} + \mathtt{RTOL(i)}\, |y_{n+1}(i)|\right),$$

or

$$\mathtt{TOL} = \max_{i=1,\cdots,m} \left(\max\left(\mathtt{ATOL(i)}, \mathtt{RTOL(i)}\, |y_{n+1}(i)|\right)\right).$$

1.5 Periodic orbits in the restricted three body problem

We illustrate the methods of this chapter with an example from Celestial Mechanics, the circular co-planar restricted three body problem (CCR3B). One considers two bodies moving, in a plane, in circular orbits about their center of mass, and a third body of negligible mass moving around in the same plane. The CCR3B problem is usually solved in rotating rectangular Cartesian coordinates with the origin at the center of mass of the primaries and the x-axis along the line of centers. The units of distance, time and mass are chosen so the separation of the primaries is the unit distance, the period of rotation is 2π and the gravitational constant is one. If μ denotes the ratio of the mass of lighter primary to the total mass, the positions of the lighter and heavier masses are $(\mu - 1, 0)$ and $(\mu, 0)$ respectively, and the equations of motion for the massless particle are then, if $(y_1, y_2, y_3, y_4) = (x, y, x', y')$,

$$y_1' = y_3, \tag{8}$$

$$y_2' = y_4, \tag{9}$$

$$y_3' = y_1 + 2y_4' - \mu' \frac{y_1 - \mu}{D_1} - \mu \frac{y_1 + \mu'}{D_2}, \tag{10}$$

$$y_4' = y_2 - 2y_3' - \mu' \frac{y_2}{D_1} - \mu \frac{y_2}{D_2}, \tag{11}$$

$$D_1 = ((y_1 - \mu)^2 + y_2^2)^{3/2}, \qquad D_2 = ((y_1 + \mu')^2 + y_2^2)^{3/2}, \tag{12}$$

$$\mu' = 1 - \mu. \tag{13}$$

The table below gives $y_1(0)$, $y_2'(0)$, and the period T for 4 periodic orbits for the CCR3B problem. The other initial conditions are always $y_1'(0) = 0$ and $y_2(0) = 0$. The orbits are symmetric about the x-axis. The values of $y_1(0)$ are exact, and $y_2'(0)$ and T are correct to 16 significant figures. The first 2 problems have $\mu = 0.012277471$ and model the Earth and Moon as the primaries. The Earth is at $(0.012277471, 0)$ and the Moon at $(-0.987722529, 0)$. The remaining 2 problems have $\mu = 0.000953875$ and model the Sun and Jupiter as the primaries. The Sun is at $(0.000953875, 0)$ and Jupiter at $(-0.999046125, 0)$.

	$y_1(0)$	$y_2'(0)$	Period
1	-0.994000E+00	0.2113898796694503E+01	0.5436795439260190E+01
2	-0.994000E+00	0.2031732629557337E+01	0.1112434033726609E+02
3	0.102745E+01	-0.4033448829049041E-01	0.1837131640001890E+03
4	0.976680E+00	0.6119162392641083E-01	0.1773324113152448E+03