**Numerical methods in mechanics**[1]

D. E. Soper[2]
University of Oregon
Physics 611, Theoretical Mechanics
22 October 2012

# 1 Introduction

In these notes we will investigate numerical methods for solving mechanics problems. In general, if we start from a lagrangian, we get equations $\ddot{q}_i = f_i(\{\dot{q}, q\}_N)$. Often, these equations will be too complicated to solve analytically. In that case, we can look for a numerical solution. When our book by Goldstein was first written, this was very difficult. Now, it is easy, so we should know how to do it.

The first thing to do is to rewrite our equations in the form

$$\begin{aligned} \dot{v}_i &= f_i(\{v, q\}_N) \;, \\ \dot{q}_i &= v_i \;. \end{aligned} \tag{1}$$

This is twice as many equations as we started with, but they are first order differential equations. With a somewhat different formulation of mechanics, we can get equations of the form

$$\begin{aligned} \dot{p}_i &= f_i(\{p, q\}_N) \;, \\ \dot{q}_i &= g_i(\{p, q\}_N) \;. \end{aligned} \tag{2}$$

Thus, we are seeking numerical solutions to first order differential equations of the form

$$\frac{d}{dt}\, y_i(t) = f_i(\{y(t)\}_n), \tag{3}$$

where there are now twice as many variables, which we call $y_i$, as we started with. We suppose that there are initial conditions,

$$y_i(0) \;\text{ are given.} \tag{4}$$


[2]soper@uoregon.edu

We can also consider the case that the equations have the form of Eq. (3) except that $f$ depends explicitly on the time $t$. We leave that small modification for later.

We will think of $y$ and $f$ as vectors in an $n$-dimensional space and mostly not write the indices $i$.

## 2 Simplest numerical solution

Let's suppose that the equations (3) are too complicated to solve analytically. Then we can get a numerical solution. The simplest way to do that is to let $\Delta t$ be a small time increment and try to get $y(t)$ approximately for $t = 0$, $t = \Delta t$, $t = 2\Delta t$, $t = 3\Delta t$, .... We could start with the given value $y(0)$ and then let

$$\tilde{y}(\Delta t) = y(0) + \Delta t\, f(\tilde{y}(0)). \tag{5}$$

I use $\tilde{y}$ for this to remind us that this is an approximation: $\tilde{y}(\Delta t)$ is not exactly $y(\Delta t)$. Then we know $\tilde{y}(\Delta t)$ and we can calculate

$$\tilde{y}(2\Delta t) = \tilde{y}(\Delta t) + \Delta t\, f(\tilde{y}(\Delta t)). \tag{6}$$

Then we know $\tilde{y}(2\Delta t)$ and we can calculate $\tilde{y}(3\Delta t)$. Conintuing in this fashion, once we know $\tilde{y}(k\Delta t)$ we can calculate $\tilde{y}((k+1)\Delta t)$ by

$$\tilde{y}((k+1)\Delta t) = \tilde{y}(k\Delta t) + \Delta t\, f(\tilde{y}(k\Delta t)). \tag{7}$$

Assuming that we want results for $0 \le t \le T$, we stop when $k\,\Delta t = T$. Then the total number of steps is $N_s = T/\Delta T$, or the nearest integer to that.

How accurate is this method? Consider the first step. In order to keep the analysis as simple as possible, we assume tht there is only one component in the vector $y$. We know that

$$y(\Delta t) = y(0) + \Delta t\, \dot{y}(0) + (\Delta t)^2 \ddot{y}(0) + \cdots\,. \tag{8}$$

or

$$y(\Delta t) = y(0) + \Delta t\, f(y(0)) + (\Delta t)^2 \ddot{y}(0) + \cdots\,. \tag{9}$$

Thus

$$y(\Delta t) - \tilde{y}(\Delta t) = (\Delta t)^2 \ddot{y}(0) + \cdots\,. \tag{10}$$

We can write this as

$$y(\Delta t) - \tilde{y}(\Delta t) = \mathcal{O}((\Delta t)^2), \tag{11}$$

meaning that the this difference vanishes at least as fast as $(\Delta t)^2$ as $\Delta t \to 0$.

Now consider that we have gotten to a time $t$ after making $t/\Delta t$ steps. At the next step, we make two kinds of errors. First, we have

$$y(t + \Delta t) = y(t) + \Delta t\, f(y(t)) + (\Delta t)^2 \ddot{y}(t)) + \cdots . \tag{12}$$

or

$$y(t + \Delta t) = y(t) + \Delta t\, f(y(t)) + \mathcal{O}((\Delta t)^2). \tag{13}$$

Here we see the error from taking a step. But we don't know $y(t)$ exactly, so we must take the step from a slightly wrong place. That is, we calculate

$$\tilde{y}(t + \Delta t) = \tilde{y}(t) + \Delta t\, f(\tilde{y}(t)). \tag{14}$$

Consider the difference between the step we take in the calculation and the actual step:

$$\mathcal{E} = [\tilde{y}(t + \Delta t) - \tilde{y}(t)] - [y(t + \Delta t) - y(t)] \tag{15}$$

We have

$$\begin{aligned} \mathcal{E} &= \Delta t\, [f(\tilde{y}(t)) - f(y(t))] + \mathcal{O}((\Delta t)^2) \\ &= \Delta t \left\{ [\tilde{y}(t) - y(t)] f'(\tilde{y}(t)) + \mathcal{O}([\tilde{y}(t) - y(t)]^2) \right\} \\ &\quad + \mathcal{O}((\Delta t)^2). \end{aligned} \tag{16}$$

Now, how far were we off at time $t$? Let us try the following hypothesis. At each step we make an additional error of order $(\Delta t)^2$. After a finite time t, we have taken $t/\Delta t$ steps, so we have made an error of order $(\Delta t)^2 \times t/\Delta t = t\, \Delta t$. Is this hypothesis self-consistent? Assuming the hypothesis, we have

$$[\tilde{y}(t) - y(t)] \sim \mathcal{O}(\Delta t). \tag{17}$$

Then

$$\begin{aligned} \mathcal{E} &= \Delta t \left\{ \mathcal{O}(\Delta t) f'(\tilde{y}(t)) + \mathcal{O}((\Delta t)^2) \right\} + \mathcal{O}((\Delta t)^2) \\ &= \mathcal{O}((\Delta t)^2) \end{aligned} \tag{18}$$

That is we continue to make errors of order $\Delta t$ in each step, so our hypothesis is self-consistent.

We conclude that the error after a finite time t is of order $\Delta t$. The error goes to zero as $\Delta t \to 0$, but not very fast.

## 3 Second order Runge-Kutta method

Now let's see how to do better. Suppose that we have an approximate value $\tilde{y}(t)$ for the vector $y$ at time $t$. To build the approximation to $y$ at time $t+\Delta t$ we define

$$k_i = \Delta t\, f_i(\tilde{y}(t)) \tag{19}$$

and put

$$\tilde{y}(t+\Delta t) = \tilde{y}(t) + \Delta t\, f_i\big(\tilde{y}(t) + \tfrac{1}{2}k\big)\,. \tag{20}$$

Iteration of this formula starting at $\tilde{y}(0) = y(0)$ gives $\tilde{y}(t)$ for $t = n\Delta t$, $n = 1, 2, \dots, T/\Delta t$.

Let us estimate the error we make in taking one step,

$$\mathcal{E}_i(t) = [\tilde{y}_i(t+\Delta t) - \tilde{y}_i(t)] - [y_i(t+\Delta t) - y_i(t)]. \tag{21}$$

We have, from the definition of the method,

$$\mathcal{E}_i(t) = \Delta t\, f_i\big(\tilde{y}(t) + \tfrac{1}{2}\Delta t\, f(\tilde{y}(t))\big) - [y_i(t+\Delta t) - y_i(t)]. \tag{22}$$

Expanding $y_i(t+\Delta t)$ about $\Delta t = 0$ we have

$$\mathcal{E}_i(t) = \Delta t\, f_i\big(\tilde{y}(t) + \tfrac{1}{2}\Delta t\, f(\tilde{y}(t))\big) - \Delta t\, \dot{y}_i(t) - \tfrac{1}{2}(\Delta t)^2 \ddot{y}_i(t) + \mathcal{O}((\Delta t)^3). \tag{23}$$

For $\dot{y}_i(t)$ we can use the differential equation to write

$$\dot{y}_i(t) = f_i(y(t)). \tag{24}$$

For $\ddot{y}_i(t)$ we can write

$$\ddot{y}_i(t) = \partial_j f_i(y(t))\ \dot{y}_j(t). \tag{25}$$

where $\partial_j = \partial/\partial y_j$ and there is an implicit sum over $j$. Again using the differential equation, this is

$$\ddot{y}_i(t) = \partial_j f_i(y(t))\ f_j(y(t)). \tag{26}$$

Then

$$\begin{aligned}\mathcal{E}_i(t) \;&=\; \Delta t\, f_i\big(\tilde{y}(t) + \tfrac{1}{2}\Delta t\, f(\tilde{y}(t))\big) \\ &\quad -\Delta t\, f_i(y(t)) - \tfrac{1}{2}(\Delta t)^2 \partial_j f_i(y(t))\ f_j(y(t)) + \mathcal{O}((\Delta t)^3).\end{aligned} \tag{27}$$

Now we need to expand $f$ in the first term about $y = \tilde{y}(t)$:

$$f_i\left(\tilde{y}(t) + \tfrac{1}{2}\Delta t\, f(\tilde{y}(t))\right) = f_i\left(\tilde{y}(t)\right) + \partial_j f_i\left(\tilde{y}(t)\right)\ \tfrac{1}{2}\Delta t\, f_j(\tilde{y}(t)) + \mathcal{O}((\Delta t)^2). \tag{28}$$

This gives

$$
\begin{aligned}
\mathcal{E}_i(t) &= \Delta t\, f_i\left(\tilde{y}(t)\right) + \tfrac{1}{2}(\Delta t)^2\, \partial_j f_i\left(\tilde{y}(t)\right)\;\; f_j(\tilde{y}(t)) \\
&\quad -\Delta t\, f_i(y(t)) - \tfrac{1}{2}(\Delta t)^2 \partial_j f_i(y(t))\;\; f_j(y(t)) + \mathcal{O}((\Delta t)^3). \qquad (29)
\end{aligned}
$$

So far, we have not used any hypothesis about the cumulative error. We can see that if the cumulative error were zero, that is, if $\tilde{y} = y$, then the order $(\Delta t)^1$ and $(\Delta t)^2$ terms would cancel and we would be left with an error $\mathcal{E}$ in the step of order $(\Delta t)^3$. But an error in each step of order $(\Delta t)^3$ would lead to a cumulative error after $T/\Delta t$ steps of order $(\Delta t)^2$.

Let us, therefore try the hypothesis that the error in each step is

$$
\mathcal{E}_i(t) = \mathcal{O}((\Delta t)^3) \qquad (30)
$$

so that cumulative error is

$$
\tilde{y}_i(t) - y_i(t) = \mathcal{O}((\Delta t)^2). \qquad (31)
$$

Then we can write

$$
\begin{aligned}
f_i(\tilde{y}(t)) &= f_i(y(t)) + \mathcal{O}((\Delta t)^2) \\
\partial_j f_i(\tilde{y}(t)) &= \partial_j f_i(y(t)) + \mathcal{O}((\Delta t)^2). \qquad (32)
\end{aligned}
$$

This gives, after cancellation,

$$
\mathcal{E}_i(t) = \mathcal{O}((\Delta t)^3), \qquad (33)
$$

consistent with our hypothesis.

We conclude that the second order Runge-Kutta method produces a cumulative error of order $(\Delta t)^2$ as $\Delta t \to 0$.

# 4 Fourth order Runge-Kutta method

Now let's see how to do even better. Suppose that we have an approximate value $\tilde{y}(t)$ for the vector $y$ at time $t$. To build the approximation to $y$ at time $t + \Delta t$ we define

$$
\begin{aligned}
k_{1,i} &= \Delta t\, f_i\left(\tilde{y}(t)\right) \\
k_{2,i} &= \Delta t\, f_i\left(\tilde{y}(t) + \tfrac{1}{2}k_1\right) \\
k_{3,i} &= \Delta t\, f_i\left(\tilde{y}(t) + \tfrac{1}{2}k_2\right) \\
k_{4,i} &= \Delta t\, f_i\left(\tilde{y}(t) + k_3\right) \qquad (34)
\end{aligned}
$$

and put

$$\tilde{y}(t+\Delta t) = \tilde{y}(t) + \frac{1}{6}\,k_1 + \frac{1}{3}\,k_2 + \frac{1}{3}\,k_3 + \frac{1}{6}\,k_4. \tag{35}$$

Iteration of this formula starting at $\tilde{y}(0) = y(0)$ gives $\tilde{y}(t)$ later times.

With this method, the error in each step is

$$\mathcal{E}_i(t) = \mathcal{O}((\Delta t)^5) \tag{36}$$

so that cumulative error is

$$\tilde{y}_i(t) - y_i(t) = \mathcal{O}((\Delta t)^4). \tag{37}$$

The argument is similar to that for the second order Runge-Kutta method, but there is quite a lot more algebra needed.

---

*Exercise 4.1:* Consider a double star system with two stars, 1 and 2, moving in circular orbit about their center of mass, which we take to be at the origin of coordinates. We choose the coordinate system so that the stars move in the $(x^1, x^2)$ plane. Then we consider the problem to be two dimensional and omit $x^3$ throughout. The stars have masses

$$m_1 = \frac{M}{1+\lambda}\,, \qquad m_2 = \frac{\lambda M}{1+\lambda}\,, \tag{38}$$

with the mass ratio $\lambda < 1$. The positions of the stars are

$$\begin{aligned} \tilde{\boldsymbol{x}}_1 &= (-\lambda R\cos(\omega t), -\lambda R\sin(\omega t))\ , \\ \tilde{\boldsymbol{x}}_2 &= (R\cos(\omega t), R\sin(\omega t))\ . \end{aligned} \tag{39}$$

Show that the total mass $M$ of the stars is related to $R$ and $\omega$ by

$$GM = (1+\lambda)^3 R^3\,\omega^2. \tag{40}$$

Now suppose that this star system contains a planet, whose mass $m$ is tiny compared to the star masses. We want to investigate the motion of the planet. For this purpose it is convenient to use a coordinate system that is rotating with angular velocity $\omega$, so that the stars now have fixed coordinates

$$\begin{aligned} \boldsymbol{x}_1 &= (-\lambda R, 0)\ , \\ \boldsymbol{x}_2 &= (R, 0)\ . \end{aligned} \tag{41}$$

Denote the position of the planet in this system by $\boldsymbol{x} = (x, y)$ and its velocity be $\boldsymbol{v} = (v_x, v_y)$. (We suppose that the initial conditions are such that $x^3 = v^3 = 0$ at time zero, so that the planet remains in the 1-2 plane.) Show that the equations of motion for the planet can be written (for motion in the 1-2 plane) as

$$\frac{d\boldsymbol{x}}{dt} = \boldsymbol{v} \tag{42}$$

and

$$\frac{d\boldsymbol{v}}{dt} = -2\,\boldsymbol{\omega} \times \boldsymbol{v} - \boldsymbol{\nabla} U(\boldsymbol{x}) \tag{43}$$

where

$$\boldsymbol{\omega} \times \boldsymbol{v} = (-\omega v_y, \omega v_x) \tag{44}$$

and

$$U = -\tfrac{1}{2}\,\omega^2(x^2 + y^2) - \frac{(1+\lambda)^2 R^3 \omega^2}{[(x + \lambda R)^2 + y^2]^{1/2}} - \frac{\lambda(1+\lambda)^2 R^3 \omega^2}{[(x - R)^2 + y^2]^{1/2}}. \tag{45}$$

Find the energy for this system. Is it conserved? Use the form of the energy to argue qualitatively that if the planet starts near to one of the stars without too much speed, it can stay in the solar system. But if it is going too fast or is not near enough to one of the stars, it may be tossed out of the solar system.

Now program this for numerical evaluation and plot some planet paths that illustrate these qualitative conclusion for the case $\lambda = 0.02$. I suggest using the 4th order Runge Kutta subroutine and choosing the unit of time so that $\omega = 1$ and the unit of distance so that $R = 1$.

Next, use your 4th order Runge Kutta subroutine to investigate what happens if the planet if it starts with a very small velocity at Lagrange point $L_2$ between the stars and what happens if it starts with a very small velocity at Lagrange point $L_4$ that has a positive $y$. You will need to find where these Lagrange points are.

Try starting the planet at the $L_2$ Lagrange point with a velocity $\boldsymbol{v} = (0.1, -0.1)$ and make a plot of the resulting orbit. Also, give a numerical result for $\boldsymbol{x}$ at time $t = 7.000$ (in the units with $\omega = 1$ and $R = 1$). (Be careful, you should check that your answer does not change much if you double the number of points. I found that I needed 10000 points in order to get a numerically stable answer.)

Then try starting the planet at the $L_4$ Lagrange point with a velocity $\boldsymbol{v} = (0.001, -0.0002)$ and make a plot of the resulting orbit. Also, give a numerical result for $\boldsymbol{x}$ at time $t = 60$.