CPSC 303: SUMMARY OF ODE NUMERICAL METHODS, STIFFNESS, AND CENTRAL FORCE PROBLEMS (AND BEYOND)

JOEL FRIEDMAN

Copyright: Copyright Joel Friedman 2024. Not to be copied, used, or revised without explicit written permission from the copyright owner.

Disclaimer: The material may sketchy and/or contain errors, which I will elaborate upon and/or correct in class. For those not in CPSC 303: use this material at your own risk...

The goal of this note is to review our discussion of ODE approximation methods, [A&G], Sections 16.2 and 16.3, and to discussion on "stiff equations," including central force problems.

Note that [A&G] solve y' = f(t, y) with the initial condition y(a) = c for greater generality. For simplicity, we solve y' = f(y) and $y(0) = y_0 \in \mathbb{R}$. When we want to use *n*-dimensional ODE's, we switch to $\mathbf{y} = \mathbf{f}(\mathbf{y})$ and $\mathbf{y}(0) = \mathbf{y}_0 \in \mathbb{R}$ (and hence $\mathbf{y} = \mathbf{y}(t)$ is a function $\mathbb{R} \to \mathbb{R}^n$); all ODE numerical methods work similarly in the *n*-dimensional case, and in the more general case of $\mathbf{y}' = \mathbf{f}(t, \mathbf{y})$.

1. BASIC STIFF EQUATIONS

Euler's method to solve $\mathbf{y}' = \mathbf{f}(\mathbf{y})$ subject to $\mathbf{y}(0) = \mathbf{y}_0$ is to pick a small h > 0and set $t_i = ih$, and use the recurrence

$$\mathbf{y}_{i+1} = \mathbf{y}_i + h\mathbf{f}(\mathbf{y}_i)$$

for $i = 0, 1, \ldots$, where \mathbf{y}_i is an approximation for $\mathbf{y}(ih)$.

One standard situation which makes this problematic is the equation y' = aywith a < 0. In this case $y(t) = e^{at}y_0$ which decays as $t \to \infty$. It is still true that for any T > 0, Euler's method to approximate y(T) converges to the true solution as $h \to 0$, since y(T) is approximated by $y_{T/h}$ (for simplicity we assume T/h is an integer), and

$$y_{T/h} = (1+ah)^{T/h} \xrightarrow{h \to 0} e^{aT} \mathbf{y}_0.$$

However, until h is roughly -1/a (when a < 0), Euler's method doesn't look qualitatively like the solution: namely, if h > -1/a, then $(1+ah)^i$ oscillates between positive and negative, and if h > -2/a, then (1 + ah) < -1, and so |1 + ah| > 1, so Euler's method gives y_i with

(1)
$$|y_i| = |1 + ah|^i |y_0|$$

which tends to infinity exponentially in i.

There are a number of remarks to make:

Research supported in part by an NSERC grant.

JOEL FRIEDMAN

- (1) With Euler's method applied to y' = ay and a < 0, one has to tailor h to the size of 1/|a| to get a numerical approximation that qualitatively resembles the true solution. The equation y' = ay is said to be "stiff" if a is a negative number of "large" absolute value (hence "stiff" is not a precise term).
- (2) Here is the real point: consider the system

(2)
$$\mathbf{y}' = A\mathbf{y}, \quad A = \operatorname{diag}(d_1, d_2) = \begin{bmatrix} d_1 & 0\\ 0 & d_2 \end{bmatrix} = \begin{bmatrix} -1000 & 0\\ 0 & 1 \end{bmatrix}$$

(so $d_1 = -1000$ and $d_2 = 1$). For the initial condition y(0) = (1, 1), we easily see that (see Section 3)

$$\mathbf{y}(t) = (y_1(t), y_2(t)) = (e^{d_1 t}, e^{d_2 t}) = (e^{-1000t}, e^t).$$

So $y_1(t)$ is insignificant for t = 1; however, if $h \ge 3/1000$, then $|1 - h(-1000)| \ge 2$, and Euler's method behaves disastrously because of $y_1(t)$ (see (1)). By contrast, h = 3/1000 is a very reasonable step size to solve $y'_2 = d_2y_2 = y_2$ and t = 1. Hence, although the $y_1(t)$ is negligible for t = 1, and you may not "physically detect $y_1(t)$ in practice," the mere existence of y_1 causes Euler's method to fail.

- (3) Similarly, in Section 3 we will explain that the same *stiffness* and can arise in solving ODE's $\mathbf{y}' = A\mathbf{y}$ when A has a very large, negative eigenvalue compared to the other eigenvalues. Hence the *stiffness* of an ODE may not be as obvious as that for the example y' = ay or y' = Ay when A is a diagonal matrix. A similar remark holds of $\mathbf{y}' = \mathbf{f}(\mathbf{y})$, where you have to linearly approximate $f(\mathbf{y})$ to detect stiffness. Hence if you are solving $\mathbf{y}' = \mathbf{f}(\mathbf{y})$ where the function \mathbf{f} is poorly understood, you may encounter stiffness that is far less obvious than (2).
- (4) Here is a remedy, at least for (2): the problem of "stiffness" can be (sometimes) be fixed by using a method such as backward Euler's method:

(3)
$$\mathbf{y}_{i+1} = \mathbf{y}_i + h\mathbf{f}(\mathbf{y}_{i+1}).$$

(we discuss this and other methods in Section 2). Note that (3) is *implicit*, in that it requires us to solve for \mathbf{y}_{i+1} .¹ Regardless, we can analyze backward Euler's method for y' = ay: we have

$$\mathbf{y}_{i+1} = \mathbf{y}_i + ha\mathbf{y}_{i+1},$$

 \mathbf{SO}

$$\mathbf{y}_{i+1} = \frac{1}{1 - ha} \mathbf{y}_i.$$

Hence for a < 0, this iterative scheme has the right qualitative behaviour.

However, backward Euler has its limitations: if a > 0 and h > 2/a, then \mathbf{y}_i oscillates in sign and is qualitatively different than the true solution.

$$\mathbf{y}_{i+1} = \mathbf{y}_i + h\mathbf{f}(\mathbf{y}_{i+1}) = \mathbf{y}_i + h\mathbf{f}(\mathbf{y}_i + h\mathbf{f}(\mathbf{y}_{i+1}))$$

and keep iterating on the rightmost \mathbf{y}_{i+1} ; equivalently, setting

$$\Psi(\mathbf{y}) = \mathbf{y}_i + h\mathbf{f}(\mathbf{y}),$$

 $\mathbf{2}$

¹Presumably, for h sufficiently small we can write

on presumes that for small enough h, the sequence $\mathbf{y}_i, \Psi(\mathbf{y}_i), \Psi^2(\mathbf{y}_i), \ldots$ will converge to a solution of $\mathbf{y} = \Psi(\mathbf{y})$, and this solution presumably is the \mathbf{y}_{i+1} we seek.

2. Forward and Backward Euler, and Higher Order ODE Numerical Schemes

Backward Euler scheme is a *first order accurate* ODE approximation, just like forward Euler. Let us review the computation (see also [A&G], page 488), and make some brief remarks on other schemes.

2.1. Review of Forward Euler and Explicit Trapezoidal. If y = y(t) is sufficiently differentiable, and t is fixed and h > 0 is small, Taylor's theorem tells us that

$$y(t+h) = y(t) + hy'(t) + O(h^2),$$

where the order h^2 term, $O(h^2)$, actually equals $\frac{h^2}{2}y''(\xi)$ for some $\xi \in (t, t + h)$. Hence the constant in $O(h^2)$ includes a second derivative bound.

Hence, for the ODE y' = f(y), subject to $y(0) = y_0$, we choose a small h, set $t_i = ih$ for i = 0, 1, 2, ..., and we have

(4)
$$y(t_{i+1}) = y(t_i) + hf(y(t_i)) + O(h^2);$$

dropping the truncation term we use the recurrence

$$y_{i+1} = y_i + hf(y_i)$$

with y_i being approximation of $y(t_i) = y(ih)$. This yields forward Euler's method. The fact that

(5)
$$d_i \stackrel{\text{def}}{=} y(t_{i+1}) - \left(y(t_i) + hf(y(t_i))\right) = O(h^2),$$

leads us to expect that the global error $e_i = y(t_i) - y_i$ will be roughly *i* times $O(h^2)$. Hence for fixed *T*, with h > 0 small, we have $t_i = T$ for i = T/h (which—for simplicity—we assume is an integer², and expect $y(T) - y_{T/h}$ to be T/h times $O(h^2)$, or roughly O(Th).

In general, for any ODE approximation method we define d_i in (5) similarly, and when $d_i = O(h^2)$ —or, equivalently, we drop a term of order $O(h^2)$ in the analog of (4), we expect a global error of O(Th) in approximating y(T) and we say the method is accurate to first order. Similarly, if the local error d_i is $O(h^q)$, we call the method accurate to (q-1)-th order.

Similarly, a longer calculation with Taylor series shows that

$$y(t+h) = y(t) + h\left(\frac{y(t) + Y(t)}{2}\right) + O(h^3), \text{ where } Y(t) = y(t) + hf(y(t)),$$

and dropping the $O(h^3)$ term (whose constant really involves third derivative bounds) we get the explicit Trapezoidal rule, a second order method, namely

$$y_{i+1} = y_i + h\left(\frac{f(y_i) + f(Y_i)}{2}\right), \text{ where } Y_i = y_i + hf(y_i),$$

where $d_i = O(h^3)$, and is therefore accurate to second order: we expect the ("global") error in approximating y(T) as $y_{T/h}$ to be $(T/h)O(h^3) = O(Th^2)$.

A warning: to get evidence that Euler's method has an O(Th) error in approximating y(T), we tested them with the simple ODE y' = ay, where we can explicitly compute the exact solution and what Euler's method yields. However, to prove the global error is really O(Th) for a general ODE y' = f(y) (or, more generally

²If T/h is not an integer, we could approximate y(T) as an appropriate convex combination of $y_{|T/h|}$ and $y_{[T/h]}$ and get similar results

 $\mathbf{y}' = \mathbf{f}(t, \mathbf{y})$, [A&G], top of page 489, shows that one needs assumptions on f(y) and y(t), and the O(Th) has some hidden constants based on these assumptions. Fortunately, these hidden constants (mostly) show up in the ODE y' = ay, which is another selling point of these simple examples.

2.2. Backward Euler and Beyond. We can similarly see that

$$y(t+h) = y(t) + hy'(t+h) + O(h^2),$$

which gives the backward Euler method

$$y_{i+1} = y_i + hf(y_{i+1});$$

since we dropped an $O(h^2)$ term, backward Euler method is first order accurate.

Another implicit scheme is based on

$$y(t+h) = y(t) + h\left(\frac{y'(t) + y'(t+h)}{2}\right) + O(h^3),$$

which gives us the second order method

$$y_{i+1} = y_i + h\left(\frac{f(y_i) + f(y_{i+1})}{2}\right).$$

In Section 16.3, [A&G] refer to this method as the *implicit Trapezoidal method*.

[A&G] describe a number of other ODE approximation methods in Section 16.3, which one can analyze similarly (although the Taylor series computations get progressively more difficult). One popular scheme is RK4, a classical, explicit *Runga-Kutta* scheme of order 4, which is somewhat reminiscent of Simpson's rule for integration.

A fuller discussion of the above could easily be an entire course; e.g., see the textbook *Numerical Methods for Ordinary Differential Equations* by John C. Butcher. Here we see that ODE methods, including RK4, can be built using the theory of directed graphs.

3. Stiffness in Eigenvalues

3.1. Stiffness in Diagonal Matrices. Say that A is the diagonal matrix

$$A = \operatorname{diag}(d_1, d_2) = \begin{bmatrix} d_1 & 0\\ 0 & d_2 \end{bmatrix}$$

Then the two-dimensional ODE $\mathbf{y}' = A\mathbf{y}$ becomes

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}' = \mathbf{y}' = A\mathbf{y} = \begin{bmatrix} d_1y_1 \\ d_2y_2 \end{bmatrix}$$

which is really two separate one-dimensional ODE's

$$y_1' = d_1 y_1, \quad y_2' = d_2 y_2.$$

So, for example, if d_2 is of moderate size (either positive or negative or zero), and $d_1 < 0$ and $|d_1|$ is much greater than $|d_2|$, then although the $y_1(t)$ term, proportional to $e^{d_1 t}$, decays much more quickly to 0 than $y_2(t)$ term, and the main behaviour of $\mathbf{y}' = A\mathbf{y}$ lies in $y_2(t)$, Euler's method will have a large oscillating term when h is a bit larger than $2/(-d_1)$.

Hence the relatively insignificant $y_1(t)$ term can make Euler's method disastrous, even when h is small enough compared to $1/|d_2|$.

Similarly remarks hold A is an $n \times n$ diagonal matrix diag (d_1, \ldots, d_n) . Note that in the case the eigenvalues of A are d_1, \ldots, d_n . (For this reason [A&G] sometimes uses the more suggestive equation $y' = \lambda y$ instead of y' = ay or $y' = d_i y$.³

3.2. Stiffness in (Similar and) Diagonalizable Matrices. The discussion about $\mathbf{y}' = A\mathbf{y}$ goes through almost the same for diagonalizable matrices; the point is that if $\mathbf{z}' = B\mathbf{z}$ is another ODE system, and A and B are similar matrices, then the exact solutions of these two ODE systems and the numerical approximation schemes for these two are essentially the same (up to similarity).

So say that A, S are $n \times n$ matrices with S invertible. Say that $\mathbf{y} = \mathbf{y}(t)$ satisfies $\mathbf{y}' = A\mathbf{y}$, and set $\mathbf{z}(t) = S\mathbf{y}(t)$, which is equivalent to $\mathbf{y}(t) = S^{-1}\mathbf{z}(t)$. Then we easily see that

$$\mathbf{z}'(t) = \left(S\mathbf{y}(t)\right)' = S\mathbf{y}'(t) = SA\mathbf{y}(t) = SA\left(S^{-1}\mathbf{z}(t)\right),$$

and hence

$$\mathbf{z}' = B\mathbf{z}$$
, where $B = SAS^{-1}$

We now show that Euler's method applied to $\mathbf{y}' = A\mathbf{y}$ with initial condition $\mathbf{y}(0) = \mathbf{y}_0$ is equivalent to Euler's method applied to $\mathbf{z}' = B\mathbf{z}$ with initial condition $\mathbf{z}(0) = \mathbf{z}_0$ where $\mathbf{z}_0 = S\mathbf{y}_0$: indeed, if

$$\mathbf{y}_{i+1} = (I + hA)\mathbf{y}_i$$

then setting $\mathbf{z}_i = S\mathbf{y}_i$ for all i we have

$$\mathbf{z}_{i+1} = S\mathbf{y}_{i+1} = S(I+hA)\mathbf{y}_i = S(I+hA)S^{-1}\mathbf{z}_i = (I+hB)\mathbf{z}_i$$

which is just Euler's method for **z**.

It follows that if A is diagonalizable, i.e., for some invertible S we have $B = SAS^{-1}$ and $B = \text{diag}(d_1, \ldots, d_n)$ is a diagonal matrix, then Euler's method applied to A is equivalent to Euler's method applied to B. Hence Euler's method has similar problems due to "stiffness" in the presence of "large negative" eigenvalues; while these large negative eigenvalues give insignificant (or quickly decaying) terms to the ODE $\mathbf{y}' = A\mathbf{y}$, these eigenvalues can ruin Euler's method.

4. Stiffness in the Central Force Problems: Complex Eigenvalues without Complex Eigenvalues

The equation $\mathbf{y}' = A\mathbf{y}$ can have more subtle stiffness issues in systems that arise when A has complex (non-real) eigenvalues. This arises in the central force problem in celestial mechanics. Here we discuss such examples without using complex numbers.

The typical example is $\mathbf{y}' = A\mathbf{y}$, where

$$A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix};$$

the eigenvalue equation for A, namely $\det(\lambda I - A) = 0$ gives $\lambda^2 + 1 = 0$, and hence we get $\lambda = \pm \sqrt{-1}$, which are "complex numbers," in fact "purely imaginary" complex numbers. It is simplest to explain the issues here without any reference to complex numbers. Let us switch to celestial mechanics notation, using for d/dt instead of '.

³Also, [A&G] also use y(a) = c as their initial condition, rather than $y(0) = y_0$.

4.1. The Harmonic Oscillator. Consider an $x: \mathbb{R} \to \mathbb{R}$ satisfying $\ddot{x} = -x$. We have encountered this ODE before, and know that its general solution is $x(t) = C_1 \sin(x) + C_2 \cos(x)$; hence this ODE is usually called "the harmonic oscillator." Hence

$$\frac{d}{dt} \begin{bmatrix} \dot{x} \\ x \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \dot{x} \\ x \end{bmatrix},$$

i.e.,

$$\dot{\mathbf{y}} = A\mathbf{y}, \text{ where } \mathbf{y} = \begin{bmatrix} \dot{x} \\ x \end{bmatrix}, A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

Defining the "energy" of this system to be

Energy(t)
$$\stackrel{\text{def}}{=} (\dot{x})^2(t) + x^2(t) = ||\mathbf{y}(t)||^2$$

(the L^2 -norm), we easily see that (d/dt)Energy(t) = 0 for all t, so the energy is independent of time t. Of course, we can regard energy as a function of $\mathbf{y} = \mathbf{y}(t)$, in that Energy(t) =Energy $(\mathbf{y}(t)) = ||\mathbf{y}||^2$ (which is independent of t).

Euler's method starting at time t = 0 and with step size h therefore gives the approximation

$$\mathbf{y}_{i+1} = (I + hA)\mathbf{y}_i,$$

and hence

$$\mathbf{y}_i = (I + hA)^i \mathbf{y}_0$$

where $y_0 = \mathbf{y}(0) = (\dot{x}(0), x(0))$, and where \mathbf{y}_i is an approximation to y(mh). Note that for any $\mathbf{z} = (z_1, z_2) \in \mathbb{R}^2$ and $h \in \mathbb{R}$ we have

$$(I+hA)\begin{bmatrix} z_1\\ z_2 \end{bmatrix} = \begin{bmatrix} z_1-hz_2\\ z_2+hz_1 \end{bmatrix},$$

and hence

$$||(I+hA)\mathbf{z}||^2 = (z_1 - hz_2)^2 + (z_2 + hz_1)^2 = (1+h^2)(z_1^2 + z_2^2) = (1+h^2)||\mathbf{z}||^2.$$

It follows that for Euler's method,

$$\|\mathbf{y}_{i+1}\|^2 = (1+h^2)\|\mathbf{y}_i\|^2,$$

and hence

$$\|\mathbf{y}_i\|^2 = (1+h^2)^i \|\mathbf{y}_0\|^2.$$

In other words, after *m* iterations of Euler's method, the approximation \mathbf{y}_i to $\mathbf{y}(ih)$ has energy that is exactly $(1 + h^2)^i$ of the original energy (which does not change in the exact solution).

This is not really so bad (not yet...): say that we want to approximate $\mathbf{y}(T)$ for some fixed T > 0; since \mathbf{y}_i approximates y(ih), we need *i* iterations of Euler's method with i = T/h (for simplicity assume this is an integer). Hence

Energy
$$(\mathbf{y}_{T/h}) = (1+h^2)^{T/h}$$
Energy $(\mathbf{y}_0),$

so the change in energy, for small h, is a factor of

$$(1+h^2)^{T/h} = \left(e^{h^2+O(h^4)}\right)^{T/h} = e^{Th+O(Th^3)}.$$

So this is what you'd expect, given that Euler's method is accurate to order one. This means that for small $\delta > 0$, if you can tolerate the energy increasing by a factor of $1 + \delta$, then it suffices to have

$$Th + O(Th^3) = \log(1+\delta) = \delta - O(\delta^2),$$

or roughly $Th \approx \delta$, i.e.,

$$h \approx \delta/T.$$

4.2. The More Oscillatory Oscillator. Consider the second order ODE $\ddot{x} = -c^2 x$ with c > 0. We have encountered this ODE before, and know that it's general solution is $x(t) = C_1 \sin(cx) + C_2 \cos(cx)$, which is therefore periodic with period $2\pi/c$; the larger c, the shorter the period and the more quickly the solution oscillates.

In this case we define the energy to be $E(t) = (\dot{x})^2(t) + c^2 x^2(t)$, and similarly show that E(t) is constant.

We may similarly write this ODE as

$$\dot{\mathbf{y}} = A\mathbf{y}, \text{ where } \mathbf{y} = \begin{bmatrix} \dot{x} \\ cx \end{bmatrix}, A = \begin{bmatrix} 0 & -c \\ c & 0 \end{bmatrix}.$$

Hence $E(t) = ||\mathbf{y}(t)||^2$. Show that after *m* steps of Euler's method with step size *h* applied to $\dot{\mathbf{y}} = A\mathbf{y}$, the energy of the approximation increases by a factor of $(1 + c^2 h^2)^m$.

Now matters are somewhat worse: if you want Euler's method to approximate y(T) to have an increase in energy no more than a factor of $1 + \delta$, then (1) ch has to be sufficiently small, and (2) in this case, you need

$$\left(1+c^2h^2\right)^{T/h} \le 1+\delta,$$

so dropping smaller order terms we have roughly speaking

$$(T/h)c^2h^2 \le \log(1+\delta),$$

and so you must take

$$h = \delta/(Tc^2).$$

Hence fixing T and δ , h must be proportional to $1/c^2$.

This may not be surprising, but this means that in systems with some less significant but more rapidly oscillating structure, the rapid oscillations may cause serious changes in energy if h is not tailored not only to δ , T, but also to c.

For another matter, if $ch = \rho$ is some fixed number, then $(1 + c^2 h^2)^i = (1 + \rho)^i$, which means that after a fixed number of steps in *i*, the energy doubles. So $h = \rho/c$ makes Euler's method disastrous after a fixed number of steps.

4.3. Trouble in the Central Force Problem. Consider the one-dimensional central force problem: $m\ddot{x} = -mu(|x|)x/|x|$, or, more simply, $\ddot{x} = -u(|x|)x/|x|$. We can rewrite this is $\ddot{x} = -c(x)x$, where c(x) = u(|x|)/|x|.

It follows that if t_0 is fixed, for t near t_0 we have that c(x(t)) is near $c_0 = c(x(t_0))$. It follows that to have the energy not increase more than a factor of $1 + \delta$ between t_0 and $t_0 + T$, we need $h = T\delta/c_0^2$.

Imagine we use Newton's law $u(r) = g/r^2$ where g is constant, and hence $c(x) = g/|x|^3$. When |x| is small, it follows that we need $h = T\delta|x|^6/g^2$ (for T small) to avoid more than a factor of $1 + \delta$ increase in numerical energy. This quantifies how small h needs to be when |x(t)| becomes small. This should confirm our intuition that if g (gravity) is increased, then the force is greater and we need a smaller step size; similarly, as |x| deceases, we need a much smaller step size.

For similar reasons, in the central force problem $m\ddot{\mathbf{x}} = -mg\mathbf{x}/||\mathbf{x}||^3$, your step size h may need to be roughly $T\delta ||x||^6/g^2$, as a "ballpark" estimate.

Note that since the exact energy in the central force problem $m\ddot{\mathbf{x}} = -mu(\|\mathbf{x}\|)\mathbf{x}/\|\mathbf{x}\|$ is

$$\frac{1}{2}m\|\dot{\mathbf{x}}\|^2 + mU(\|\mathbf{x}\|)$$

where U'(r) = u(r), i.e., $\int u(r) dr = U(r)$, when c(r) = u(r)/r is roughly constant for r near r_0 , then u(r) is roughly cr and U(r) is roughly $(1/2)(c^2r^2)$ for r near r_0 . Hence the true energy is roughly m/2 times $\|\mathbf{y}\|^2$ with \mathbf{y} as in the previous sections.

EXERCISES

- (1) One exercise.
- (2) More exercises to follow.

DEPARTMENT OF COMPUTER SCIENCE, UNIVERSITY OF BRITISH COLUMBIA, VANCOUVER, BC V6T 1Z4, CANADA.

E-mail address: jf@cs.ubc.ca *URL*: http://www.cs.ubc.ca/~jf

8