Math 317 Assignment 3

Due in class: November 17th, 2016

Instructions: Submit a hard copy of your solution with your name and student number. (**No name = zero grade!**) You must include all relevant program code, electronic output and explanations of your results. Write your own codes and comment them. Late assignment will not be graded and will receive a grade of zero.

1. Consider the integral $I(f) = \int_0^1 f(x)dx$.

   (a) (5 marks) Determine the two point Gauss quadrature for I on $[0, 1]$. What is its degree of accuracy?

   Solution: By the change of variable, $x = \frac{1-0}{2}t + \frac{1+0}{2} = \frac{1+t}{2}$, the two point Gauss quadrature on $[0, 1]$ is

   $$I_{Gauss}(f) = \frac{1}{2}\left(f\left(\frac{1}{2}\left(1 - \frac{1}{\sqrt{3}}\right)\right) + f\left(\frac{1}{2}\left(1 + \frac{1}{\sqrt{3}}\right)\right)\right),$$

   with degree of accuracy of 3.

   (b) (5 marks) Determine c_0, c_1 and x_1 so that the quadrature $I_h(f) = c_0 f(0) + c_1 f(x_1)$ has the highest degree of accuracy possible. State this degree.

   Solution: Denote the quadrature as $I_h(f) := c_0 f(0) + c_1 f(x_1)$. The degree of accuracy conditions imposes

   $$\begin{cases} I(1) = I_h(1) \\ I(x) = I_h(x) \\ I(x^2) = I_h(x^2) \end{cases} \iff \begin{cases} 1 = c_0 + c_1 \\ \frac{1}{2} = c_1 x_1 \\ \frac{1}{3} = c_1 x_1^2 \end{cases} \iff \begin{cases} x_1 = \frac{2}{3} \\ c_1 = \frac{3}{4} \\ c_0 = \frac{1}{4} \end{cases}$$

   Thus

   $$I_h(f) := \frac{1}{4}\left(f(0) + 3f(\frac{2}{3})\right)$$

   Moreover, $I_h(f)$ has degree of accuracy of 2, since

   $$\frac{1}{4} = I(x^3) \neq I_h(x^3) = \frac{1}{4}\left(0 + 3\frac{8}{27}\right) = \frac{2}{9}$$

   (c) (5 marks) Recall from probability, the Gaussian distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$ is $f(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}$. In this case, compare the approximate values for I using the two point Gauss quadrature and your quadrature in part (b). (The exact value of I is $0.3413447\ldots$).

   Solution: As expected, Gauss quadrature is more accurate than the quadrature from part (b):

   $$\begin{aligned} I_{Gauss}(f) &= \frac{1}{2\sqrt{2\pi}}\left(\exp\left(-\frac{1}{2}\left(\frac{1}{2}\left(1 - \frac{1}{\sqrt{3}}\right)^2\right)\right) + \exp\left(-\frac{1}{2}\left(\frac{1}{2}\left(1 + \frac{1}{\sqrt{3}}\right)^2\right)\right)\right) \\ &\approx 0.34122114\ldots \\ I_h(f) &= \frac{1}{4\sqrt{2\pi}}\left(\exp(0) + 3\exp\left(-\frac{1}{2}\left(\frac{2}{3}\right)^2\right)\right) \approx 0.33932157\ldots \end{aligned}$$

2. Consider the integral $I(f) = \int_0^1 f(x)dx$.

(a) (8 marks) Derive the formula for the composite trapezoidal rule and its error.

Solution: Let $x_k = x_0 + kh$ where $h = \frac{b-a}{n}$. On each $[x_k, x_{k+1}]$, there is some $\xi_k \in (x_k, x_{k+1})$,

$$\int_{x_k}^{x_{k+1}} f(x)\, dx = h\left(f(x_k) + f(x_{k+1})\right) - f^{(2)}(\xi_k)\frac{h^3}{12}$$

Hence

$$\begin{aligned} I(f) &= h\sum_{k=0}^{n-1}\left(f(x_k) + f(x_{k+1})\right) + \frac{h^3}{12}\sum_{k=0}^{n-1} f^{(2)}(\xi_k) \\ &= h\sum_{k=0}^{n-1}\left(f(x_k) + f(x_{k+1})\right) + \frac{b-a}{12} f^{(2)}(\xi)h^2 \end{aligned}$$

since

$$\frac{1}{n}\sum_{k=0}^{n-1} f^{(2)}(\xi_k) = f^{(2)}(\xi)$$

for some $\xi \in (a, b)$ by the intermediate value theorem.

(b) (10 marks) Using Richard's extrapolation method, we can use the composite trapezoidal rule to derive a more accurate quadrature. Such quadrature is given by

$$I_h^R(f) = \frac{4I_{\frac{h}{2}}(f) - I_h(f)}{3}$$

where I_h denotes here the composite trapezoidal rule and has error $O(h^4)$. The goal of this question is to perform a convergence analysis of the composite Trapezoidal rule and the improved quadrature for $I(f) = \int_0^1 e^{-x}\, dx$. In order to do so write a program to approximate using both quadratures for $h = 1, 2^{-1}, \ldots 2^{-8}$, plot log(error) versus log(h) confirm their convergence rate by estimating the slopes of the lines in the loglog plot.

Solution: For small h, we expect error $\approx Ch^p$ where $p = 2$ for the composite trapezoidal rule and $p = 4$ for the Richard's quadrature. Then

$$\begin{aligned} \text{error} &\approx Ch^p \\ \log(\text{error}) &\approx p\log(h) + \log(C) \end{aligned}$$

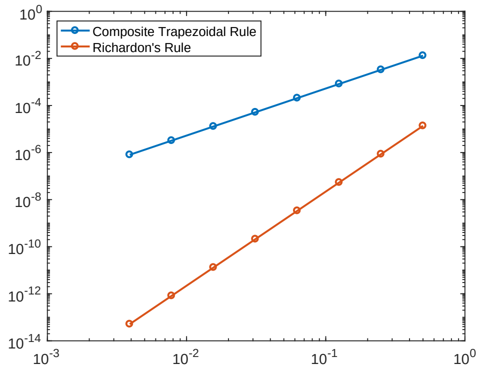
For sufficiently small h, the plot of log(error) versus log(h) should be a line with slope p. Moreover, we can estimate the slope p by

$$p \approx \frac{\log(\text{error}_{i+1}) - \log(\text{error}_i)}{\log(h_{i+1}) - \log(h_i)}$$

where $\text{error}_i = I(f) - I_{h_i}(f)$ and $h_i = 2^{-i}$. The estimated slopes are display in the next table.

h	Composite Trapezoidal Rule		Richardson's quadrature	
2^{-1}	1.311×10^{-2}	-	1.362×10^{-5}	-
2^{-2}	3.289×10^{-3}	1.996	8.558×10^{-7}	3.992
2^{-3}	8.229×10^{-4}	1.999	5.356×10^{-8}	3.998
2^{-4}	2.058×10^{-4}	2.000	3.349×10^{-9}	3.999
2^{-5}	5.144×10^{-5}	2.000	2.093×10^{-10}	4.000
2^{-6}	1.286×10^{-5}	2.000	1.308×10^{-11}	4.000
2^{-7}	3.215×10^{-6}	2.000	8.172×10^{-13}	4.001
2^{-8}	8.038×10^{-7}	2.000	5.118×10^{-14}	3.997

Errors and estimated order of convergence for both quadratures.



Loglog plot of the error versus h for both quadratures.

3. This exercise is to derive the order conditions for linear multistep method. Recall a k-step linear multi-step method for first-order initial value problems has the form:

$$\Phi_h := y_{n+1} + \sum_{i=0}^{k-1} a_i y_{n-k+1+i} - h \sum_{i=0}^{k} b_i f_{n-k+1+i} = 0,$$

where a_i, b_i are constants and $f_i := f(t_i, y_i)$.

(a) (5 marks) Denoting $a_k = 1$ and y as the exact solution to the initial value problem, show that the local truncation error is,

$$\tau_h(t_n) = \sum_{i=0}^{k} a_i y(t_{n-k+1} + ih) - h \sum_{i=0}^{k} b_i y'(t_{n-k+1} + ih).$$

Solution: By taking $a_k = 1$, we have that

$$y_{n+1} = a_k y_{n+1} = a_k y_{n-k+1-k}$$

and so

$$\Phi_h = \sum_{i=0}^{k} a_i y_{n-k+1+i} - h \sum_{i=0}^{k} b_i f_{n-k+1+i}.$$

Since y is the exact solution to the initial value problem, we have that for $i = 0, \ldots, k$,

$$f(t_{n-k+1} + ih, y(t_{n-k+1} + ih)) = y'(t_{n-k+1} + ih).$$

Hence

$$\tau_h(t_n) = \sum_{i=0}^{k} a_i y(t_{n-k+1} + ih) - h \sum_{i=0}^{k} b_i y'(t_{n-k+1} + ih).$$

(b) (5 marks) By Taylor expanding y, y' around t_{n-k+1}, show that for some $\xi_i, \eta_i \in [t_{n-k+1}, t_{n-k+1} + ih]$

$$\tau_h(t_n) = \sum_{i=0}^{k} a_i \left(\sum_{q=0}^{p} \frac{(ih)^q}{q!} y^{(q)}(t_{n-k+1}) + \frac{(ih)^{p+1}}{(p+1)!} y^{(p+1)}(\xi_i) \right) - h \sum_{i=0}^{k} b_i \left(\sum_{q=1}^{p} \frac{(ih)^{q-1}}{(q-1)!} y^{(q)}(t_{n-k+1}) + \frac{(ih)^p}{p!} y^{(p+1)}(\eta_i) \right).$$

Solution: By Taylor expanding y around t_{n-k+1}, we have for $i = 0, \ldots, k$

$$y(t_{n-k+1} + ih) = \sum_{q=0}^{p} \frac{(ih)^q}{q!} y^{(q)}(t_{n-k+1}) + \frac{(ih)^{p+1}}{(p+1)!} y^{(p+1)}(\xi_i)$$

for some $\xi_i \in [t_{n-k+1}, t_{n-k+1} + ih]$.
By Taylor expanding y' around t_{n-k+1}, we have for $i = 0, \ldots, k$

$$y'(t_{n-k+1}) = \sum_{q=1}^{p} \frac{(ih)^{q-1}}{(q-1)!} y^{(q)}(t_{n-k+1}) + \frac{(ih)^p}{p!} y^{(p+1)}(\eta_i)$$

for some $\eta_i \in [t_{n-k+1}, t_{n-k+1} + ih]$.
Plugging the above two expressions in the formula derived for $\tau_h(t_n)$ in a) leads to

$$\begin{aligned}\tau_h(t_n) = & \sum_{i=0}^{k} a_i \left(\sum_{q=0}^{p} \frac{(ih)^q}{q!} y^{(q)}(t_{n-k+1}) + \frac{(ih)^{p+1}}{(p+1)!} y^{(p+1)}(\xi_i) \right) \\ & - h \sum_{i=0}^{k} b_i \left(\sum_{q=1}^{p} \frac{(ih)^{q-1}}{(q-1)!} y^{(q)}(t_{n-k+1}) + \frac{(ih)^p}{p!} y^{(p+1)}(\eta_i) \right),\end{aligned}$$

as desired.

(c) (5 marks) Show that the local truncation error can be written in the form,

$$\tau_h(t_n) = \sum_{q=0}^{p} \left(\frac{h^q}{q!} y^{(q)}(t_{n-k+1}) C_q \right) + \frac{h^{p+1}}{(p+1)!} D,$$

where

$$C_q = \sum_{i=0}^{k} i^q a_i - q \sum_{i=0}^{k} i^{q-1} b_i, \quad D = \sum_{i=0}^{k} \left(i^{p+1} a_i y^{(p+1)}(\xi_i) - (p+1) i^p b_i y^{(p+1)}(\eta_i) \right).$$

Solution: It's enough to see that

$$\begin{aligned}& \sum_{i=0}^{k} a_i \left(\sum_{q=0}^{p} \frac{(ih)^q}{q!} y^{(q)}(t_{n-k+1}) + \frac{(ih)^{p+1}}{(p+1)!} y^{(p+1)}(\xi_i) \right) \\ = & \sum_{q=0}^{p} \frac{h^q}{q!} y^{(q)}(t_{n-k+1}) \left(\sum_{i=0}^{k} i^q a_i \right) + \frac{h^{p+1}}{(p+1)!} \sum_{i=0}^{k} i^{p+1} a_i y^{(p+1)}(\xi_i)\end{aligned}$$

and

$$
\begin{aligned}
& h \sum_{i=0}^{k} b_i \left(\sum_{q=1}^{p} \frac{(ih)^{q-1}}{(q-1)!} y^{(q)}(t_{n-k+1}) + \frac{(ih)^p}{p!} y^{(p+1)}(\eta_i) \right) \\
&= \sum_{q=1}^{p} \frac{h^q}{(q-1)!} y^{(q)}(t_{n-k+1}) \left(\sum_{i=0}^{k} i^{q-1} b_i \right) + \frac{h^{p+1}}{p!} \sum_{i=0}^{k} i^p b_i y^{(p+1)}(\eta_i) \\
&= \sum_{q=0}^{p} \frac{h^q}{q!} y^{(q)}(t_{n-k+1}) \left(q \sum_{i=0}^{k} i^{q-1} b_i \right) + \frac{h^{p+1}}{(p+1)!} \sum_{i=0}^{k} (p+1) i^p b_i y^{(p+1)}(\eta_i)
\end{aligned}
$$

(d) (3 marks) Conclude that a k-step linear multi-step method is of order p if and only if a_i, b_i satisfies $C_q = 0$ for all $q = 0, \dots, p$. Or equivalently, a_i, b_i satisfies for all $q = 0, \dots, p$,

$$
q \sum_{i=0}^{k} i^{q-1} b_i = k^q + \sum_{i=0}^{k-1} i^q a_i. \text{ (i.e. order conditions)}
$$

Solution: A k-step linear multi-step method is of order p if and only if the lowest power of h in the local truncation error is $p+1$, i.e., if and only if $C_q = 0$ for all $q = 0, \dots, p$.

4. The implicit 2-step Milne-Simpson method is:

$$
y_{n+1} = y_{n-1} + \frac{h}{3}(f_{n+1} + 4f_n + f_{n-1}).
$$

(a) (10 marks) Show that the local truncation error is $O(h^5)$.

Solution: There's two ways to solve this: use Taylor expansion or the order conditions to show that the method is of order 4 (recall that the order of a method is always the order of the local truncations error minus 1). Here we provide both proofs.

Taylor expansion approach: The local truncation error is given by

$$
\tau_h(t_n) = y(t_{n+1}) - y(t_{n-1}) - \frac{1}{3} h f(t_{n+1}, y(t_{n+1})) - \frac{4}{3} h f(t_n, y(t_n)) - \frac{1}{3} h f(t_{n-1}, y(t_{n-1}))
$$

for $n = 1, 2 \dots, N-1$. The idea is to Taylor expand every term in the above expression which is not evaluated at $t = t_n$. We

$$
y(t_{n\pm+1}) = y(t_n) \pm h y'(t_n) + \frac{h^2}{2} y''(t_n) \pm \frac{h^3}{6} y^{(3)}(t_n) + \frac{h^4}{24} y^{(4)}(t_n) + \mathcal{O}(h^5), \quad f(t_n, y(t_n)) = y'(t_n),
$$

and

$$
f(t_{n\pm 1}, y(t_{n\pm 1})) = y'(t_{n\pm 1}) = y'(t_n) \pm h y''(t_n) + \frac{h^2}{2} y^{(3)}(t_n) \pm \frac{h^3}{6} y^{(4)}(t_n) + \mathcal{O}(h^4).
$$

Plugging in the above formulas into the expression of $\tau_{n+1}(h)$ and simplifying shows that $\tau_h(t_n) = \mathcal{O}(h^5)$ as desired.

Order conditions approach: We have

$$a_1 = 0, \quad a_0 = -1, \quad b_2 = \frac{1}{3}, \quad b_1 = \frac{4}{3}, \quad b_0 = \frac{1}{3}.$$

We have to check that

$$q \sum_{i=0}^{k} i^{q-1} = k^q + \sum_{i=0}^{k-1} i^q a_i$$

for $q = 0, \ldots, 4$ and that it fails for $q = 5$. Indeed, we have

- $q = 0$

$$q \sum_{i=0}^{k} i^{q-1} = 0$$
$$k^q + \sum_{i=0}^{k-1} i^q a_i = 2^0 + 0^0 \times (-1) + 1^0 \times 0 = 0$$

- $q = 1$

$$q \sum_{i=0}^{k} i^{q-1} = 1 \left(0^0 \times \frac{1}{3} + 1^0 \times \frac{4}{3} + 2^0 \times \frac{1}{3} \right) = 2$$
$$k^q + \sum_{i=0}^{k-1} i^q a_i = 2^1 + 0^1 \times (-1) + 1^1 \times 0 = 2$$

- $q = 2$

$$q \sum_{i=0}^{k} i^{q-1} = 2 \left(0^1 \times \frac{1}{3} + 1^1 \times \frac{4}{3} + 2^1 \times \frac{1}{3} \right) = 4$$
$$k^q + \sum_{i=0}^{k-1} i^q a_i = 2^2 + 0^2 \times (-1) + 1^2 \times 0 = 4$$

- $q = 3$

$$q \sum_{i=0}^{k} i^{q-1} = 3 \left(0^2 \times \frac{1}{3} + 1^2 \times \frac{4}{3} + 2^2 \times \frac{1}{3} \right) = 8$$
$$k^q + \sum_{i=0}^{k-1} i^q a_i = 2^3 + 0^3 \times (-1) + 1^3 \times 0 = 8$$

- $q = 4$

$$q \sum_{i=0}^{k} i^{q-1} = 4 \left(0^3 \times \frac{1}{3} + 1^3 \times \frac{4}{3} + 2^3 \times \frac{1}{3} \right) = 16$$
$$k^q + \sum_{i=0}^{k-1} i^q a_i = 2^4 + 0^4 \times (-1) + 1^4 \times 0 = 16$$

- $q = 5$

$$q \sum_{i=0}^{k} i^{q-1} = 5 \left(0^4 \times \frac{1}{3} + 1^4 \times \frac{4}{3} + 2^4 \times \frac{1}{3} \right) = \frac{100}{3}$$
$$k^q + \sum_{i=0}^{k-1} i^q a_i = 2^5 + 0^5 \times (-1) + 1^5 \times 0 = 32$$

(b) (5 marks) Show that the method is zero-stable and conclude that it is convergent.

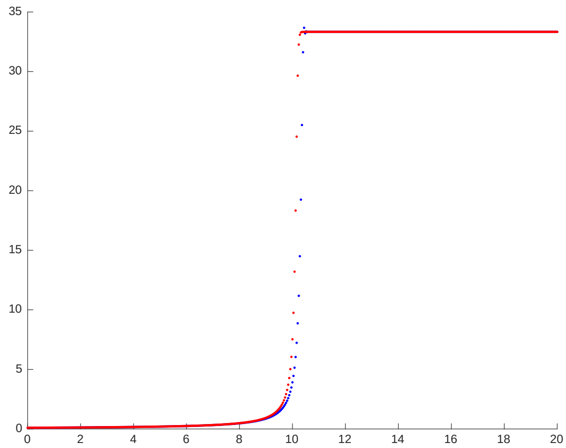
Solution: The characteristic polynomial is given by $p(\lambda) = \lambda^2 - 1$ which has roots -1 and 1. Therefore the method is zero stable. In part a), we saw that the method has order 4 and therefore it is consistent. We can then conclude that it is convergent.

5. Consider the I.V.P. on $t \in [0, T]$:

$$y' = y^2(1 - \epsilon y).$$

(a) (10 marks) For $T = 20, y(0) = 0.1, \epsilon = 0.03$, use the forward Euler and Trapezoidal method to solve the I.V.P. with $N = 500$ and plot both solutions versus t.

Solution: See Figure.



Plot of the solution versus t with $N = 500$ for the Euler method (blue) and the Trapezoidal method (red).

(b) (4 marks) For the equilibrium solution $y^* = 1/\epsilon$, show that the I.V.P. is approximately,

$$y' \approx -\frac{1}{\epsilon}(y - y^*) \text{ when } |y - y^*| \text{ is small.}$$

Hint: Taylor expand $f(y) = y^2(1 - \epsilon y)$ around $y = y^*$.

Solution: We have

$$f'(y) = -y^2\epsilon + 2y(1 - y\epsilon)$$

and so

$$f'(y^*) = -\frac{1}{\epsilon}.$$

Hence we have

$$\begin{aligned} f(y) &= f(y^*) + f'(y^*)(y - y^*) + \mathcal{O}(|y - y^*|^2) \\ &\approx -\frac{1}{\epsilon}(y - y^*) \end{aligned}$$

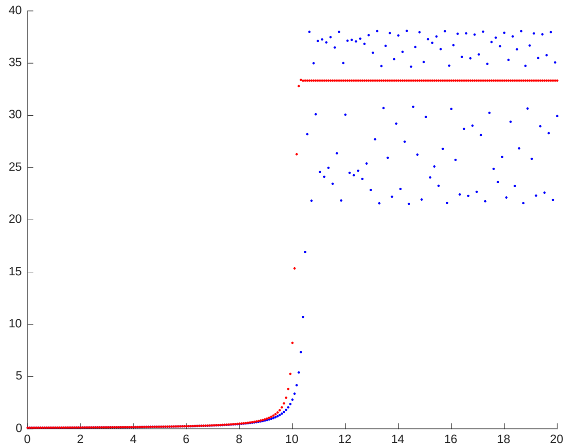
when $|y - y^*|$ is small.

(c) (4 marks) Plot both solutions when $N = 250$ and use part (b) to explain what is happening.

Solution: See Figure. From the plot it's clear the Euler method didn't converge. The reason for this is that, as shown in *b*) the problem behaves like a stiff problem with $\lambda = -\frac{1}{\epsilon}$ and so for the Euler method to converge we need

$$h < -\frac{2}{\lambda} = 2\epsilon = 0.06.$$

However, with $N = 250$, h does not satisfy the above condition since $h = 0.08$. Note as well, that in part *a*), since $N = 500$, $h = 0.04$, thus explaining why the Euler method behaved better.



Plot of the solution versus t with $N = 250$ for the Euler method (blue) and the Trapezoidal method (red).

6. Consider the non-dimensionalized pendulum problem

$$\begin{cases} \theta''(t) + \sin(\theta(t)) = 0, & t \in [0, T], \\ \theta(0) = a, \\ \theta'(0) = b. \end{cases}$$

Let $\theta(t)$ denote the exact solution.

(a) (2 marks) Write the second order equation as a system of first order equations.

Solution: Let $y_1 = \theta$ and $y_2 = \theta'$. Hence the second order equation can be rewritten as

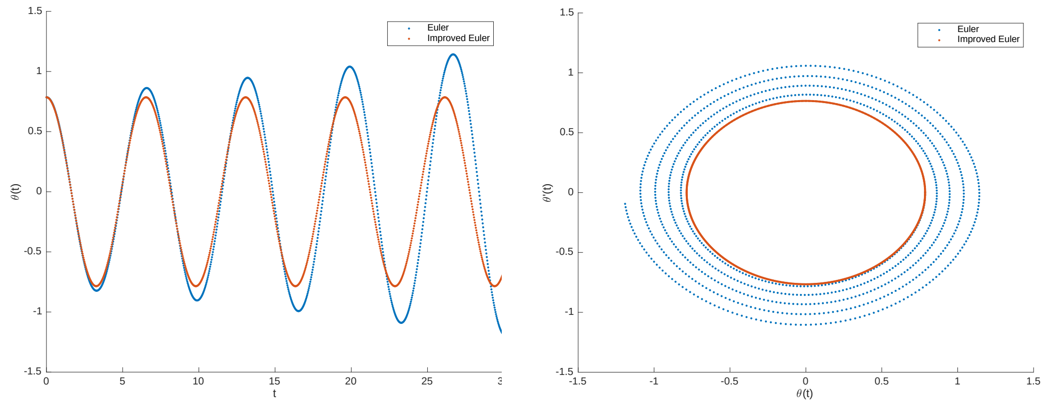
$$\mathbf{y}'(t) = \begin{bmatrix} y_1'(t) \\ y_2'(t) \end{bmatrix} = \begin{bmatrix} \theta'(t) \\ \theta''(t) \end{bmatrix} = \begin{bmatrix} \theta'(t) \\ -\sin(\theta(t)) \end{bmatrix} = \begin{bmatrix} y_2(t) \\ -\sin(y_1(t)) \end{bmatrix} = \mathbf{F}(t, \mathbf{y}(t))$$

with

$$\mathbf{y}(0) = \begin{bmatrix} y_1(0) \\ y_2(0) \end{bmatrix} = \begin{bmatrix} a \\ b \end{bmatrix} = \mathbf{y}_0$$

(b) (10 marks) For $N = 1000$, use the forward Euler and improved Euler's method to solve the first order system for $a = \pi/4, b = 0$ up to $T = 30$. Plot the two solutions as $\theta(t)$ versus t and as $\theta'(t)$ versus $\theta(t)$.

Solution: See Figure.



Plot the two solutions as $\theta(t)$ versus t and as $\theta'(t)$ versus θ.

(c) (2 marks) Let $E(t) = \frac{(\theta'(t))^2}{2} - \cos(\theta(t))$ denote the energy of the pendulum. Show that the energy is conserved, i.e.,

$$\frac{d}{dt}E(t) = 0.$$

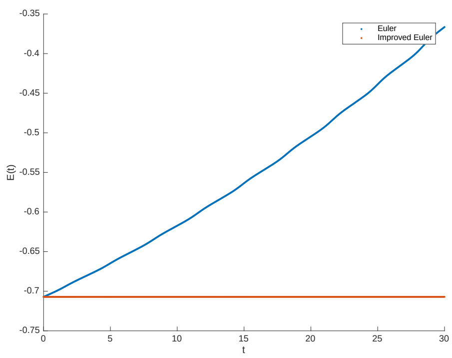
Solution: We have

$$\frac{d}{dt}\left(\frac{(\theta')^2}{2} - \cos(\theta)\right) = \theta''\theta' + \theta' \sin(\theta) = \theta'(\theta'' + \sin(\theta)) = 0,$$

where in the last equation we used the fact that θ is the exact solution of the pendulum problem.

(d) (2 marks) Plot the energy computed using the two methods as a function of t. Which method has the least "energy drift"?

Solution: See Figure. The method that has the least "energy drift" is the improved Euler; the energy remains approximately constant.



Plot the two solutions as θ versus t and as θ' versus θ.