CPSC 403/542
Assignment 1 - Solutions and beyond

1.

$$\begin{aligned} x' &= -y \\ y' &= x \end{aligned}$$

We observe that the forward Euler solution spirals out, the backward Euler solution spirals in and the trapezoidal solution is just right.

Explanation: the equation determining the circle is

$$x^2 + y^2 = r^2$$

and this should remain (at least close to) invariant under discretization for the curve to close, i.e. we want at each step $x_n^2 + y_n^2 = x_{n-1}^2 + y_{n-1}^2$.

Forward Euler:

$$\begin{aligned} x_n &= x_{n-1} - h y_{n-1} \\ y_n &= y_{n-1} + h x_{n-1} \end{aligned}$$

Squaring each equation and adding, we get

$$x_n^2 + y_n^2 = (1 + h^2)(x_{n-1}^2 + y_{n-1}^2) > x_{n-1}^2 + y_{n-1}^2$$

making it clear why the solution curve spirals out.

Backward Euler:

$$\begin{aligned} x_n &= x_{n-1} - h y_n \Rightarrow x_n + h y_n = x_{n-1} \\ y_n &= y_{n-1} + h x_n \Rightarrow y_n - h x_n = y_{n-1} \end{aligned}$$

Squaring each equation and adding, we get

$$x_n^2 + y_n^2 = (1 + h^2)^{-1}(x_{n-1}^2 + y_{n-1}^2) < x_{n-1}^2 + y_{n-1}^2$$

making it clear why the solution curve spirals in.

Trapezoidal (or midpoint):

$$\begin{aligned} x_n &= x_{n-1} - h/2(y_{n-1} + y_n) \Rightarrow x_n + h/2\ y_n = x_{n-1} - h/2\ y_{n-1} \\ y_n &= y_{n-1} + h/2(x_{n-1} + x_n) \Rightarrow y_n - h/2\ x_n = y_{n-1} + h/2\ x_{n-1} \end{aligned}$$

Squaring each equation and adding, we get

$$(1 + h^2/4)(x_n^2 + y_n^2) = (1 + h^2/4)(x_{n-1}^2 + y_{n-1}^2)$$

Hence the quadratic invariant is preserved: $x_n^2+y_n^2 = x_{n-1}^2+y_{n-1}^2$, and the circle closes well.

Beyond Q. 1

In view of Q. 3 below, the sucess of the trapezoidal scheme for drawing circles seems to be explainable by the forward- and backward- Euler errors cancelling each other out. Another possibility that suggests itself then is a *composite Euler* scheme which applies backward Euler to the first equation and forward Euler to the second:

$$\begin{aligned} x_n &= x_{n-1} - hy_n \\ y_n &= y_{n-1} + hx_{n-1} \end{aligned}$$

Note that the second equation is used to calculate y_n first, and then the first equation is used to calculate x_n in an explicit manner.

It is easy to check that this scheme *does not* preserve the invariant (which is the circle's equation). Still, it will produce a curve that looks, when plotted, like a circle for fairly coarse h (and certainly for $h = .02$). The reason is that this scheme is *symplectic*. A long discussion of symplectic maps and methods can be found in Hairer-Norsett-Wanner. In this simple case it boils down to the fact that the scheme is area preserving, i.e. if we consider a bunch of trajectories starting from a whole set of initial conditions then the exact flow preserves the area of this set in time, and so does the composite Euler method. In contrast, the forward and backward Euler methods expand and shrink the area of the initial-value set, respectively.

2. (a) Symmetry is simple:

$$\psi(y_{n-1}, y_n, h) = f(t_{n-1}+h/2, \frac{y_{n-1}+y_n}{2}) = f(t_n-h/2, \frac{y_n+y_{n-1}}{2}) = \psi(y_n, y_{n-1}, -h)$$

For a constant coefficient ODE $\mathbf{y}' = A\mathbf{y}$, the midpoint scheme is the same as the trapezoidal scheme:

$$h^{-1}(\mathbf{y}_n - \mathbf{y}_{n-1}) = A\frac{1}{2}(\mathbf{y}_n + \mathbf{y}_{n-1}) = \frac{1}{2}(A\mathbf{y}_n + A\mathbf{y}_{n-1})$$

The A-stability of the midpoint scheme therefore follows from what we showed for the trapezoidal scheme in class.

To show second order it is a good idea to expand Taylor series about the midpoint: denote $\mathbf{y} = \mathbf{y}(t_{n-1/2})$ and similarly for derivatives. Then

$$\frac{\mathbf{y}(t_n) + \mathbf{y}(t_{n-1})}{2} = y + \frac{h^2}{8}\mathbf{y}''' + \cdots$$

$$\frac{\mathbf{y}(t_n)-\mathbf{y}(t_{n-1})}{h}-\mathbf{f}(\frac{\mathbf{y}(t_n)+\mathbf{y}(t_{n-1})}{2})=\mathbf{y}'+\frac{h^2}{24}\mathbf{y}'''-\mathbf{f}(\mathbf{y}+\frac{h^2}{8}\mathbf{y}'''+\cdots)+\cdots=O(h^2)$$

(b) For the variable coefficient test equation $y'=\lambda(t)y$, the midpoint scheme yields

$$y_n=\frac{2+h\lambda(t_{n-1/2})}{2-h\lambda(t_{n-1/2})}y_{n-1}$$

The argument that

$$|\frac{2+h\lambda(t_{n-1/2})}{2-h\lambda(t_{n-1/2})}|\le 1$$

whenever $\mathcal{R}\mathrm{e}(\lambda)\le 0$ is not different from the constant coefficient case. For the trapezoidal scheme, on the other hand,

$$y_n=\frac{2+h\lambda(t_{n-1})}{2-h\lambda(t_n)}y_{n-1}$$

The factor $|\frac{2+h\lambda(t_{n-1})}{2-h\lambda(t_n)}|$ is no longer guaranteed to be below 1. In fact, if $\lambda(t)$ is real and negative and $h\lambda(t_{n-1})<h\lambda(t_n)\ll -2$ then clearly

$$|\frac{2+h\lambda(t_{n-1})}{2-h\lambda(t_n)}|\approx|\frac{\lambda(t_{n-1})}{\lambda(t_n)}|>1$$

3. (a) Half a step of forward Euler gives

$$y_{n-1/2}=y_{n-1}+h/2\ f(y_{n-1})$$

Following this by half a step of backward Euler gives

$$y_n=y_{n-1/2}+h/2\ f(y_n)$$

Adding these two equations up clearly yields the trapezoidal step.

(b) Half a step of backward Euler first gives

$$y_{n-1/2}=y_{n-1}+h/2\ f(y_{n-1/2})$$

Then forward Euler gives

$$y_n=y_{n-1/2}+h/2\ f(y_{n-1/2})$$

Adding these equations up gives

$$y_n=y_{n-1}+h\ f(y_{n-1/2})$$

Subtracting these equations from one another confirms that indeed

$$y_{n-1/2}=(y_n+y_{n-1})/2$$

so the midpoint step is obtained.

(c) Since each trapezoidal step is half a forward-Euler step (denote f) followed by half a backward-Euler step (denote b), N trapezoidal steps give the sequence $(fb)(fb)(fb)(fb)(fb)(fb)\ldots(fb)$. Writing this sequence as f followed by the rest $(bf)(bf)(bf)(bf)(b\ldots(bf)$ and one more b, we obtain the required result.

Another way: Writing the trapezoidal scheme down,

$$y_n = y_{n-1} + h/2(f_n + f_{n-1}), \quad n = 1, 2, \ldots N$$

and summing these expressions up from $n = 1$ to $n = N$ yields

$$y_N - y_0 = h/2\ f_0 + \ h \sum_{n=1}^{N-1} f_n \ + h/2\ f_N$$

On the other hand, doing half a forward-Euler step, $y_{1/2} = y_0 + h/2 f_0$, followed by $N-1$ midpoint steps from one midstep to the next,

$$y_{n+1/2} = y_{n-1/2} + h f_n, \quad n = 1, \ldots, N-1$$

and then a half backward-Euler step $Y_N = Y_{N-1/2} + h/2 f_N$ and summing all these up gives the same expression for $y_N - y_0$.

(d) An example can be derived from the previous exercise, part (b). Consider the scalar problem

$$y' = 10^{18}(t-1)y, \quad 0 < t < 1 - 10^{-12}$$

and $y(0) = 10^{-6}$. Use a uniform step $h = 10^{-4}$. The exact solution and the midpoint solution both remain below 10^{-6} in magnitude, so they are close to each other by less than 10^{-5}. For the trapezoidal scheme, on the other hand, we have $(t-1)10^{18}h < -10^6 h = -10^2 \ll -2$, so

$$y_N = \frac{2 + h\lambda(t_{N-1})}{2 - h\lambda(t_N)} y_{N-1} \approx -\frac{\lambda(t_{N-1})}{\lambda(t_N)} y_{N-1} \approx \ldots \approx (-1)^N \frac{\lambda(t_0)}{\lambda(t_N)} y_0$$

So

$$|y_N| \approx 10^{12-6} = 10^6$$

4. The obtained table is given below.

Observations: If we keep $h = b/n$ fixed and vary b then the discretization remains the same – this is only a stretching transformation – but the integration is for a longer time, so we are measuring the accumulation of local errors. If we keep b fixed and vary h, we can observe the order of the method.

Midpoint and trapezoidal schemes: As b is kept fixed and h is varied, their 2nd order accuracy is clearly reflected in the table. As h is kept fixed and b is varied, no (siginificant) error accumulation occurs.

b	N	forward Euler	backward Euler	trapezoidal	midpoint
1	10	.35e-1	.36e-1	.29e-2	.22e-2
	20	.18e-1	.18e-1	.61e-3	.51e-3
10	100	.39	.45	.41e-2	.26e-2
	200	.20	.22	.10e-2	.66e-3
100	1000	2.46	25.90	.42e-2	.26e-2
	2000	1.88	6.07	.10e-2	.66e-3
1000	1000	2.72	5.9e+301	.41	.30
	10000	2.72	1.79e+11	.42e-2	.26e-2
	20000	2.72	6.70e+5	.10e-2	.66e-3
	100000	2.49	29.77	.42e-4	.26e-4

Table 0.1: Maximum errors for long interval integration of $y' = (\cos t)y$

Forward Euler: For b not large, 1st order accuracy is observed when varying h (e.g. $b = 10$). For b large the error is almost constant arount $e \approx 2.72$. Not much improvement is observed when h is reduced.

Backward Euler: For b not large, 1st order accuracy is observed when varying h (e.g. $b = 10$). For b large the error behaves badly near $h = 1$ (in fact the denominator $1 - h\cos(t_n)$ may hit 0 when $h = 1$, which causes the scheme to blow up). The error then reduces quickly when h is reduced, but it remains large.

Explanations beyond Q. 4

The behaviour for $b \leq 10$ is as expected. The interesting part is when b is large and h is small such that bh is large. In this case the error may be large, in general.

For our problem the exact solution is $y(t) = e^{\sin t}$. It is periodic with period 2π. For simplicity let

$$b = 2\pi K$$

where K is a positive integer, and consider forward Euler, backward Euler and midpoint with stepsize $h = \frac{2\pi}{\nu}$ for some even integer ν.

Forward Euler:

$$y_n = (1 + h\cos t_{n-1})y_{n-1} = \ldots = \pi_{j=0}^{n-1}(1 + h\cos t_j) = e^{\sum_{j=0}^{n-1} \ln(1+h\cos t_j)}$$

Over one period we have

$$y_\nu = e^{\sum_{j=0}^{\nu-1} \ln(1+h\cos t_j)}$$

Taylor's expansion gives

$$\ln(1 + h\cos t) = 0 + h\cos t - \frac{1}{2}h^2\cos^2 t + \frac{1}{3}h^3\cos^3 t - \ldots$$

Now,

$$\sum_{j=0}^{\nu-1} \cos^l t_j = 0, \quad \forall l \text{ odd}$$

so

$$\sum_{j=0}^{\nu-1} \ln(1 + h\cos t_j) = -[\frac{h^2}{2}\sum_{j=0}^{\nu-1}\cos^2 t_j + \frac{h^4}{4}\sum_{j=0}^{\nu-1}\cos^4 t_j + \ldots] \geq -Jh$$

where J is some positive constant. For any n after i periods ($i \leq K$) we therefore have

$$y_n = e^{\sum_{j=0}^{n-1}\ln(1+h\cos t_j)} \leq e^{-iJh+O(1)}$$

Thus $y_n \to 0$ as $i \to \infty$ and, since the exact solution equals e at its maximum in each period, the maximum error tends to $e - 0 = e$.

Backward Euler:

$$y_n = (1 - h\cos t_{n-1})^{-1} y_{n-1} = \ldots = \pi_{j=1}^{n}(1 - h\cos t_j)^{-1} = e^{-\sum_{j=1}^{n}\ln(1-h\cos t_j)}$$

Of course, if $h = 1$ then backward Euler hits a singuarity at $t = \pi$ and blows up.

For h smaller, an analogous argument to the above yields

$$-\sum_{j=1}^{\nu}\ln(1 - h\cos t_j) \geq Jh$$

so, after i periods

$$y_n \geq e^{iJh+O(1)}$$

and we get $y_n \to \infty$ as $i \to \infty$.

Midpoint:

$$y_n = \frac{1 + h\cos t_{n-1/2}}{1 - h\cos t_{n-1/2}} y_{n-1} = \ldots = \pi_{j=1}^{n}\frac{1 + h\cos t_{j-1/2}}{1 - h\cos t_{j-1/2}} = e^{\sum_{j=1}^{n}\ln(1+h\cos t_{j-1/2}) - \ln(1-h\cos t_{j-1/2})}$$

where we let $h = h/2$ to avoid dragging the factor 2 about. Now, as before expand to get

$$\ln(1 + h\cos t) - \ln(1 - h\cos t) = 2h\cos t + 4h^3\cos^3 t + \ldots$$

which is an infinite sum in only the odd powers of cos. However, we saw before that as we sum these terms over a period they each sum to 0. Therefore, there is no error accumulation for this special problem. In general we get an $O(h^2)$ global truncation error corresponding to the action within one period, regardless of how many periods are integrated.