Solutions to Study Questions

Numerical Methods For Differential Equations, *FMNN10*

Leonid Pototskiy

January 18, 2023

Use the links below to navigate.

Contents

Special thanks to

- Erik Danielsson, π20
- Oskar Watsfeldt, ₣20
- Lennart Küssner, ₣20
- David Stålmarck, π20
- Erica Fritz, ₣20

for finding mistakes and providing solutions to some problems.

1 Initial Value Problems I

1.1. The four prinicples are

1. Discretization,
2. Polynomials and linear algebra,
3. Iteration,
4. Linearization.

1.2. Let $x = \theta, y = \theta'$. Then we can separate the equations as follows:

$$\begin{cases} x' = y & , x(0) = \theta_0, \\ y' = -\frac{g}{L}\sin x & , y(0) = \theta_0'. \end{cases}$$

1.3. No, the equations are non-linear because of the products $y_1(t) \cdot y_2(t)$.

1.4. The θ-method for $\dot{y} = f(t, y)$ is defined as

$$y_{n+1} = y_n + h(\theta f(t_{n+1}, y_{n+1}) + (1-\theta) f(t_n, y_n)).$$

The table below demonstrates which θ-values correspond to the standard methods.

θ	**Method**
0	Explicit Euler
0.5	Trapezoidal Rule
1	Implicit Euler

Table 1: The standard methods are special cases of the θ-method.

1.5. The local error ℓ_{n+1} is defined as

$$\ell_{n+1} = \hat{y}_{n+1} - y(t_{n+1}),$$

where $\hat{y}_{n+1}$ is one iteration of the given method (and problem) with initial value $y(t_n)$. Here, $y(\cdot)$ refers to the actual solution.

1.6. The global error e_{n+1} is defined as

$$e_{n+1} = y_{n+1} - y(t_{n+1}).$$

1.7. A time stepping method is said to be convergent if for every fixed $T = N \cdot h$, all $n \leq N$ and a given $h > 0$ we have

$$\lim_{N\to\infty} ||y_{n,h} - y(t_n)|| = 0,$$

where $y_{n,h}$ is the time stepping method's n'th iteration with step size h.

1.8. The following conditions are equivalent:

- The method is stable and the global error satisfies $||e_n|| = \mathcal{O}(h^p)$.
- The method is stable and the local error satisfies $||\ell_n|| = \mathcal{O}(h^{p+1})$.
- The order of convergence is p.

- The method is exact for polynomials of degree $\leq p$, i.e. the error is *always* zero.

For the explicit and implicit Euler methods we have $p = 1$ and for trapezoidal rule we have $p = 2$.

1.9. Using the last statement in the above question and the fact that the trapezoidal rule has $p = 2$ we try with a second and third-degree polynomial. With $P(t) = at^2 + bt + c$ we get

$$\begin{aligned} \text{LHS} &= P(t_{n+1}) - P(t_n) = a(t_{n+1}^2 - t_n^2) + b(t_{n+1} - t_n), \\ \text{RHS} &= \frac{t_{n+1} - t_n}{2}(\dot{P}(t_{n+1} + \dot{P}(t_n)) \\ &= \frac{t_{n+1} - t_n}{2}(2at_{n+1} + b + 2at_n + b) \\ &= (t_{n+1} - t_n)(a(t_{n+1} + t_n) + b) \\ &= a(t_{n+1}^2 - t_n^2) + b(t_{n+1} - t_n). \end{aligned}$$

However, for $P(t) = t^3$ we get

$$\begin{aligned} \text{LHS} &= t_{n+1}^3 - t_n^3, \\ \text{RHS} &= \frac{t_{n+1} - t_n}{2}(3t_{n+1}^2 + 3t_n^2) \\ &= \frac{3}{2}(t_{n+1} - t_n)(t_{n+1}^2 + t_n^2). \end{aligned}$$

Obviously, the two sides are not the same and thus the method is not exact for polynomials of degree 3. We've therefore verified the method is convergent with order $p = 2$.

1.10. Since $\dot{y}(t_i) = f(t_i, y(t_i))$, we get

$$f(t_i, y_i) \approx \frac{y(t_{i+1}) - y(t_{i-1})}{2h} \iff y(t_{i+1}) \approx y(t_{i-1}) + 2hf(t_i, y(t_i)).$$

1.11. The stability region is defined as the set of all $h\lambda \in \mathbb{C}$ such that $||y_n||$ is bounded for all n when the method is applied to the linear test equation $\dot{y} = \lambda y$.

Applying this to the explicit Euler method yields

$$y_{n+1} = y_n + h\lambda y_n \iff y_{n+1} = (1 + h\lambda) \cdot y_n.$$

The solution to this equation with respect to y_i is bounded iff $|1 + h\lambda| \leq 1$. Thus, the stability region is

$$D = \{z : |1 + z| \leq 1\},$$

which is a unit circle in $\mathbb{C}$ centered at $z = -1$.

1.12. We do the same thing as above and get

$$y_{n+1} = y_n + h\lambda y_{n+1} \iff y_{n+1} = \frac{1}{1 - h\lambda} \cdot y_n.$$

The stability region is therefore

$$\frac{1}{|1 - h\lambda|} \leq 1 \iff |1 - h\lambda| \geq 1 \implies D = \{z : |1 - z| \geq 1\},$$

which is the complement to a unit circle centered at $z = 1$.

1.13. This time we get

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

Here we get a stability region

$$\frac{|1 + h\lambda/2|}{|1 - h\lambda/2|} \leq 1 \iff |1 + h\lambda/2| \leq |1 + h\lambda/2|,$$

which geometrically means ”all points that lie further away from $z = 1$ than $z = -1$”. This is exactly the same as all points for which $\mathrm{Re}(z) \leq 0$. Thus,

$$D = \{z : \mathrm{Re}(z) \leq 0\}.$$

1.14. We begin by computing the eigenvalues. For A_1, A_2 it is trivial, since they are diagonal. We have

$$\begin{aligned} A_1 &: \lambda_1 = 10, \quad \lambda_2 = 1, \\ A_2 &: \lambda_1 = 1, \quad \lambda_2 = -10. \end{aligned}$$

For A_3 we find the characteristic polynomial

$$\det(A_3 - \lambda I) = \begin{vmatrix} -1-\lambda & 2 \\ 0 & 1-\lambda \end{vmatrix} = \lambda^2 - 1,$$

which has roots $\lambda_1 = 1, \lambda_2 = -1$.

For the Euclidean norm (same as the operator 2-norm), the calculations for A_1, A_2 are once again trivial. For diagonal matrices, the operator 2-norm is simply the largest absolute value of a diagonal element, i.e.

$$||A_1||_2 = ||A_2||_2 = 10.$$

For A_3 we have to do more work.

$$A_3^H A_3 = B = \begin{pmatrix} 1 & -2 \\ -2 & 5 \end{pmatrix} \implies \det(B - \lambda I) = \begin{vmatrix} 1-\lambda & -2 \\ -2 & 5-\lambda \end{vmatrix} = \lambda^2 - 6\lambda + 1,$$

which has solutions $\lambda_1 = 3 + \sqrt{8}, \lambda_2 = 3 - \sqrt{8}$. The norm is precisely the square root of the largest of these two, i.e.

$$||A_3||_2 = \sqrt{\lambda_1} = \sqrt{3 + \sqrt{8}}.$$

For the logarithmic 2-norm $\mu_2(A)$, we use the fact that it can be computed by finding the largest eigenvalue of the matrix $B = (A + A^H)/2$. Again, the computations for A_1, A_2 are trivial, since $B = A_i$ for these matrices. Thus,

$$\mu_2(A_1) = 10, \qquad \mu_2(A_2) = 1.$$

For A_3 we get

$$B = \begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix} \implies \det(B - \lambda I) = \lambda^2 - 2,$$

which yields the eigenvalues $\lambda_1 = \sqrt{2}, \lambda_2 = -\sqrt{2}$. Therefore $\mu_2(A_3) = \sqrt{2}$.

1.15. The linear test equation is given by

$$\begin{cases} \dot{y}(t) = \lambda y(t), \\ y(0) = 1, \end{cases}$$

which has the solution $y(t) = e^{\lambda t}$.

1.16. A method is A-stable if its stability region fully contains $\mathbb{C}^- = \{z : \mathrm{Re}(z) \leq 0\}$.

1.17. Let's try it with the explicit Euler method.

$$\begin{aligned}
y_{n+1} &= y_n + hAy_n \\
y_{n+1} &= y_n + hT\Lambda T^{-1}y_n \\
\underbrace{T^{-1}y_{n+1}}_{z_{n+1}} &= \underbrace{T^{-1}y_n}_{z_n} + h\Lambda T^{-1}y_n \\
z_{n+1} &= z_n + h\Lambda z_n
\end{aligned}$$

We see that we get precisely what we would get if we applied the explicit Euler method to the diagonalized system $\dot{z} = \Lambda z$.

2 Initial Value Problems II

2.1. Yes, since any consistent method (which ERK is) is also convergent.

2.2. See below.

0	0	0	0
1/3	1/3	0	0
2/3	0	2/3	0
	1/4	0	3/4

Table 2: Butcher tableau for study question 2.

2.3. We see that this is a explicit method, since A is lower triangular.

$$\begin{aligned}
Y_1' &= f(t_n, y_n), \\
Y_2' &= f(t_n + h/2, y_n + hY_1'/2), \\
Y_3' &= f(t_n + h/2, y_n + hY_2'/2), \\
Y_4' &= f(t_n + h, y_n + hY_3'), \\
y_{n+1} &= y_n + \frac{h}{6}(Y_1' + 2Y_2' + 2Y_3' + Y_4').
\end{aligned}$$

2.4. Use the equations above with $f(t_n, y_n) = \lambda y_n$ to get

$$\begin{aligned}
Y_1' &= \lambda y_n, \\
Y_2' &= \lambda y_n(1 + h\lambda/2), \\
Y_3' &= \lambda y_n(1 + h\lambda/2 + (h\lambda)^2/4), \\
Y_4' &= \lambda y_n(1 + h\lambda + (h\lambda)^2/2 + (h\lambda)^3/4).
\end{aligned}$$

This yields

$$\begin{aligned}
y_{n+1} &= y_n + \frac{h\lambda}{6}y_n\left[1 + 2(1 + h\lambda/2) + 2(1 + h\lambda/2 + (h\lambda)^2/4) + (1 + h\lambda + (h\lambda)^2/2 + (h\lambda)^3/4)\right] \\
&= (1 + h\lambda + (h\lambda)^2/2 + (h\lambda)^3/6 + (h\lambda)^4/24)y_n \iff \\
P(h\lambda) &= \sum_{k=0}^{4} \frac{(h\lambda)^k}{k!} \approx e^{h\lambda},
\end{aligned}$$

which is not surprising, since $e^{\lambda t}$ is the solution to the equation.

2.5. From the Butcher tableu we get the stage derivatives

$$\begin{aligned} Y_1' &= f(t_n + h/3, y_n + hY_1'/3) &&\iff Y_1' = \frac{1}{1 - h\lambda/3} \cdot \lambda y_n, \\ Y_2' &= f(t_n + 2h/3, y_n + hY_1'/3 + hY_2'/3) &&\iff Y_2' = \frac{1}{(1 - h\lambda/3)^2} \cdot \lambda y_n. \end{aligned}$$

Inserting this into the update equation gives

$$\begin{aligned} y_{n+1} = y_n + \frac{h}{2}(Y_1' + Y_2') = y_n \cdot \frac{1 + h\lambda/3 - (h\lambda)^2/18}{(1 - h\lambda/3)^2} &\iff \\ R(z) = \frac{1 + z/3 - z^2/18}{(1 - z/3)^2}. \end{aligned}$$

To check if the method (works for RK-methods in general) is A-stable, the stability function has to satisfy

1. All poles satisfy $\text{Re}(z) > 0$.
2. $|R(i\omega)| \leq 1 \, \forall \omega \in \mathbb{R}$.

The first condition is satisfied since the only pole is $z = 1/3$. For the second condition we have

$$R(i\omega) = \frac{1 + i\omega/3 + \omega^2/18}{(1 - i\omega/3)^2} \implies |R(i\omega)|^2 = \frac{(1 + \omega^2/18)^2 + \omega^2/9}{(1 - \omega^2/9)^2 + 4\omega^2/9} = \frac{\omega^4/18^2 + 2\omega^2/9 + 1}{\omega^4/81 + 2\omega^2/9 + 1} \leq 1 \, \forall \omega \in \mathbb{R}.$$

We conclude that this method is A-stable.

2.6. An embedded method computes two approximations simultaneously by utilizing common stage derivatives.

2.7. No ERK-method is A-stable because their stability function is a polynomial (the second condition in the exercise above cannot be satisfied). According to Dahlquist's second barrier theorem, the highest order of a A-stable multistep method is $p = 2$. Therefore no such method of order 3 exists.

2.8. RK is a special case of the more general multistep method, which in is on the form

$$y_{n+1} = \Phi(f, h, y_0, y_1, \ldots, y_n, y_{n+1})$$

for some function Φ.

2.9. A linear multistep method is zero-stable if the its generating polynomial $\rho(w)$ satisfies the *root condition*, which means that all roots of ρ lie inside (or on) the unit circle.

a) $y_{n+2} = y_{n+1}$ yields

$$\rho(w) = w^2 - w \implies w_1 = 0, w_2 = 1 \implies \text{zero-stable.}$$

b) $y_{n+2} = y_n$ yields

$$\rho(w) = w^2 - 1 \implies w_1 = 1, w_2 = -1 \implies \text{zero-stable.}$$

c) $y_{n+2} = \frac{4}{3}y_{n+1} - \frac{1}{3}y_n$ yields

$$\rho(w) = w^2 - \frac{4}{3}w + \frac{1}{3} \implies w_1 = \frac{1}{3}, w_2 = 1 \implies \text{zero-stable}$$

d) $y_{n+2} = 3y_{n+1} - 2y_n$ yields

$$\rho(w) = w^2 - 3w + 2 \implies w_1 = 1, w_2 = 2 \implies \text{not zero-stable.}$$

2.10. This is the AM4-method. It is implicit since f_{n+3} is present.

2.11. We use the theorem

$$\sum_{j=0}^{k} a_j j^m = m \sum_{j=0}^{k} b_j j^{m-1}, \qquad m = 0, 1, \ldots, p.$$

Here, interpret $0^0 = 1$. We have $k = 2, a_0 = -1, a_1 = 0, a_2 = 1, b_0 = 1/3, b_1 = 4/3, b_2 = 1/3$. From that we get

$$a_0 \cdot 0^m + a_1 \cdot 1^m + a_2 \cdot 2^m = m(b_0 \cdot 0^{m-1} + b_1 \cdot 1^{m-1} + b_2 \cdot 2^{m-1}) \iff$$
$$-0^m + 2^m = \frac{m}{3}(0^{m-1} + 4 + 2^{m-1})$$

For $m = 0$ both sides are equal to 0, for $m = 1$, both sides are equal to 2, for $m = 2$ both sides are 4, for $m = 3$ both sides are 8 and for $m = 4$ both sides are 16. For $m = 5$ however, the left hand side is 32, but the right is $\frac{100}{3}$. Thus, the method is of consistency order 4.

2.12. A k-step BDF-method is defined as

$$\sum_{j=1}^{k} \frac{\nabla^j}{j} \cdot y_{n+k} = hf(t_{n+k}, y_{n+k}) \implies \text{BDF3: } \nabla y_{n+3} + \frac{\nabla^2}{2} y_{n+3} + \frac{\nabla^3}{3} y_{n+3} = hf(t_{n+3}, y_{n+3}).$$

Here, the operator ∇^j is defined recursively as

$$\begin{cases} \nabla^j y_n = \nabla^{j-1} y_n - \nabla^{j-1} y_{n-1}, & j > 1, \\ \nabla y_n = y_n - y_{n-1}, & j = 1. \end{cases}$$

From this, we get

$$\begin{aligned} \nabla y_{n+3} &= y_{n+3} - y_{n+2}, \\ \nabla^2 y_{n+3} &= y_{n+3} - 2y_{n+2} + y_{n+1}, \\ \nabla^3 y_{n+3} &= y_{n+3} - 3y_{n+2} + 3y_{n+1} - y_n. \end{aligned}$$

The formula then becomes

$$\left(1 + \frac{1}{2} + \frac{1}{3}\right) y_{n+3} + (-1 - 1 - 1) y_{n+2} + \left(\frac{1}{2} + 1\right) y_{n+1} - \frac{1}{3} y_n = hf(t_{n+3}, y_{n+3}) \iff$$
$$\frac{11}{6} y_{n+3} - 3y_{n+2} + \frac{3}{2} y_{n+1} - \frac{1}{3} y_n = hf(t_{n+3}, y_{n+3}).$$

This method is convergent with order $p = 3$. To check zero-stability, we check if the generating polynomial fulfills the root condition.

$$\rho(w) = \frac{11}{6} w^3 - 3w^2 + \frac{3}{2} w - \frac{1}{3} \implies w_1 = 1, w_{2,3} = \frac{1}{22}(7 \pm i\sqrt{39}).$$

We have $|w_{2,3}|^2 = \frac{1}{22^2}(7^2 + 39) = \frac{2}{11} < 1$, so the method is zero-stable.

3 Boundary Value Problems I

3.1. With $N = 4$ we get the step size $h = 1/(N+1) = 0.2$. The boundary conditions are homogeneous, so the boundary condition vector is zero. We get

$$\boldsymbol{F}(\boldsymbol{y}) = \frac{1}{0.04} \cdot \underbrace{\begin{pmatrix} -2 & 1 & 0 & 0 \\ 1 & -2 & 1 & 0 \\ 0 & 1 & -2 & 1 \\ 0 & 0 & 1 & -2 \end{pmatrix}}_{T} \underbrace{\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix}}_{\boldsymbol{y}} - \underbrace{\begin{pmatrix} 0.04 + y_1^2 \\ 0.16 + y_2^2 \\ 0.36 + y_3^2 \\ 0.64 + y_4^2 \end{pmatrix}}_{\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})} = \boldsymbol{0}.$$

3.2. The Jacobian is

$$\frac{\partial \boldsymbol{F}}{\partial \boldsymbol{y}} = \begin{pmatrix} \frac{\partial f_1}{\partial y_1} & \frac{\partial f_1}{\partial y_2} & \frac{\partial f_1}{\partial y_3} & \frac{\partial f_1}{\partial y_4} \\ \frac{\partial f_2}{\partial y_1} & \frac{\partial f_2}{\partial y_2} & \frac{\partial f_2}{\partial y_3} & \frac{\partial f_2}{\partial y_4} \\ \frac{\partial f_2}{\partial y_1} & \frac{\partial f_2}{\partial y_2} & \frac{\partial f_2}{\partial y_3} & \frac{\partial f_2}{\partial y_4} \\ \frac{\partial f_4}{\partial y_1} & \frac{\partial f_4}{\partial y_2} & \frac{\partial f_4}{\partial y_3} & \frac{\partial f_4}{\partial y_4} \end{pmatrix} = \begin{pmatrix} -2y_1 - 50 & 25 & 0 & 0 \\ 25 & -2y_2 - 50 & 25 & 0 \\ 0 & 25 & -2y_3 - 50 & 25 \\ 0 & 0 & 25 & -2y_3 - 50 \end{pmatrix}.$$

3.3. The aim of the Newton method is to find roots to some function $\boldsymbol{f}(\boldsymbol{x})$. It is done iteratively in the following way:

1. Linearize $\boldsymbol{f}$ around $\boldsymbol{x}_n$.
2. The root of the linearized function is $\boldsymbol{x}_{n+1}$.

The first step results in

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}) \approx \boldsymbol{f}(\boldsymbol{x}_n) + \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}(\boldsymbol{x}_n) \cdot (\boldsymbol{x} - \boldsymbol{x}_n).$$

Then, finding the root gives $\boldsymbol{y} = \boldsymbol{0}, \boldsymbol{x} = \boldsymbol{x}_{n+1}$ and we get

$$\boldsymbol{0} = \boldsymbol{f}(\boldsymbol{x}_n) + \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}(\boldsymbol{x}_n) \cdot (\boldsymbol{x}_{n+1} - \boldsymbol{x}_n) \iff \boldsymbol{x}_{n+1} = \boldsymbol{x}_n - \left(\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}(\boldsymbol{x}_n)\right)^{-1} \cdot \boldsymbol{f}(\boldsymbol{x}_n).$$

In our case, simply replacing $\boldsymbol{f}(\boldsymbol{x})$ with $\boldsymbol{F}(\boldsymbol{y})$ yields

$$\boldsymbol{y}_{n+1} = \boldsymbol{y}_n - \left(\frac{\partial \boldsymbol{F}}{\partial \boldsymbol{y}}(\boldsymbol{y}_n)\right)^{-1} \cdot \boldsymbol{F}(\boldsymbol{y}_n).$$

3.4. There are a few different approaches that handle the Neumann-condition which result in second order convergence. The simplest is by letting

$$y'(1) \approx \frac{y_{N+1} - y_{N-1}}{2h} = \beta$$

and choosing the step size $h = 1/N$ instead of $h = 1/(N+1)$ so that $x_N = 1$ and $x_{N+1} = x_N + h$. In our case (with $N = 4 \implies h = 0.25$ and $\beta = 0$), this results in $y_5 = y_3$. Therefore, the last second derivative approximation is reduced to

$$\frac{y_5 - 2y_4 + y_3}{h^2} = \frac{2y_3 - 2y_4}{h^2}.$$

The equation $\boldsymbol{F}(\boldsymbol{y}) = \boldsymbol{0}$ becomes

$$\boldsymbol{F}(\boldsymbol{y}) = \frac{1}{0.0625} \cdot \begin{pmatrix} -2 & 1 & 0 & 0 \\ 1 & -2 & 1 & 0 \\ 0 & 1 & -2 & 1 \\ 0 & 0 & 2 & -2 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} - \begin{pmatrix} 0.0625 + y_1^2 \\ 0.25 + y_2^2 \\ 0.5625 + y_3^2 \\ 1 + y_4^2 \end{pmatrix} = \boldsymbol{0}.$$

Note the 2 instead of a 1 in the last row which resulted from the last second derivative approximation.

3.5. The two most common (and equivalent) definitions are

$$\mu[A] = \sup_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\mathrm{Re}(\langle \boldsymbol{x}, A\boldsymbol{x}\rangle)}{\langle \boldsymbol{x}, \boldsymbol{x}\rangle} = \lim_{h\to 0^+} \frac{||I+hA||-1}{h},$$

for some scalar product $\langle\cdot,\cdot\rangle$ and norm $||\boldsymbol{x}||^2 = \langle \boldsymbol{x}, \boldsymbol{x}\rangle$.

3.6. A basic property of the logarithmic norm says that

$$||A^{-1}||_\infty = -\frac{1}{\mu_\infty[A]}.$$

3.7. Another basic property of the logarithmic norm is that $\mu[A] \leq ||A||$. Thus, the first bound is sharper.

3.8. We solve this and the next exercise using the definition

$$\mu[A] = \lim_{h\to 0^+} \frac{||I+hA||-1}{h} \implies \mu[\lambda] = \lim_{h\to 0^+} \frac{|I+h\lambda|-1}{h}.$$

Let $f(h) = |1+h\lambda|$ and $\lambda = a+bi$. Then

$$\begin{aligned} f(h) &= \sqrt{(1+ha)^2+(hb)^2} && \implies f(0)=1, \\ f'(h) &= \frac{1}{2f(h)}(2a(1+ha)+2b(hb)) && \implies f'(0)=a. \end{aligned}$$

Taylor expansion yields $f(h) = 1+ah+\mathcal{O}(h^2)$ and we get

$$\mu[\lambda] = \lim_{h\to 0^+} \frac{(1+ah+\mathcal{O}(h^2))-1}{h} = \lim_{h\to 0^+}(a+\mathcal{O}(h)) = a = \mathrm{Re}(\lambda).$$

3.9. See previous exercise.

3.10. A bound is *sharp* when equality can be attained. Since the test equation has the solution $y(t) = e^{\lambda t}$, we get (again letting $\lambda = a+bi$)

$$|y(t)| = |e^{\lambda t}| = |e^{at}\cdot e^{ibt}| = e^{at} = e^{\mu[\lambda]t}.$$

In other words, the bound with the logarithmic norm is always sharp for the test equation. For the bound with the norm we get

$$e^{|\lambda|t} = e^{\sqrt{a^2+b^2}t} \geq e^{at},$$

with equality when $b=0$ and $a \geq 0$, i.e. when λ is a positive real number.

4 Boundary Value Problems II

4.1. This problem is incorrect. In fact, the method of order 0. The culprit is the discretization of the differential operator. Use the shorthands $y(x_n) = y, y'(x_n) = y', y''(x_n) = y''$ and do the Taylor expansion

$$\begin{aligned} y(x_{n+1}) &= y + hy + \frac{h^2}{2}y'' + \mathcal{O}(h^3), \\ y(x_{n-1}) &= y - hy + \frac{h^2}{2}y'' + \mathcal{O}(h^3). \end{aligned}$$

Also, note that p_n is just the function p evaluated at the grid points, i.e. $p_n = p(x_n)$. Inserting this into the differential operator discretization formula

$$\frac{p_{n+1}y_{n+1} - 2p_n y_n + p_{n-1}y_{n-1}}{h^2}$$

gives us

$$\underbrace{\frac{p_{n+1} - 2p_n + p_{n-1}}{h^2}}_{\mathcal{O}(h^2)} \cdot y + \underbrace{\frac{p_{n+1} - p_{n-1}}{2h}}_{\mathcal{O}(h^2)} \cdot 2y' + \underbrace{\frac{p_{n+1} + p_{n-1}}{2h}}_{\mathcal{O}(h^{-1})} \cdot \underbrace{hy''}_{\mathcal{O}(h)} + \mathcal{O}(h) = \mathcal{O}(1).$$

The operator discretization is therefore only of order 0. As further evidence, Erik Danielsson did an implementation of this method to verify this, see A.

4.2. The matrices are presented in the order symmetric ($A = A^T$), skew-symmetric ($A = -A^T$) and lower triangular ($i < j \implies a_{ij} = 0$).

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 2 & 1 & 0 \\ 4 & 3 & 2 & 1 \end{pmatrix}.$$

4.3. The characteristic equation has the solutions

$$6r^2 - 5r + 1 = 0 \iff r_1 = \frac{1}{2}, \quad r_2 = \frac{1}{3}.$$

Thus, the general solution is

$$u_j = c_1 2^{-j} + c_2 3^{-j}$$

for some c_1, c_2. We use the initial conditions to determine the constants.

$$\begin{cases} c_1 + c_2 = 1, \\ c_1 2^{-N-1} + c_2 3^{-N-1} = 0 \end{cases} \iff c_1 = -\frac{(2/3)^{N+1}}{1-(2/3)^{N+1}}, \quad c_2 = \frac{1}{1-(2/3)^{N+1}}.$$

The solution is therefore

$$u_j = \frac{1}{1-(2/3)^{N+1}}(3^{-j} - (2/3)^{N+1} \cdot 2^{-j}).$$

4.4. This false. A counterexample is

$$S = \begin{pmatrix} 1 & 2 \\ 1 & 1 \end{pmatrix}, \quad T = \begin{pmatrix} 5 & 1 \\ 2 & 4 \end{pmatrix}$$

for which $\lambda[S] = 1 \pm \sqrt{2}$ and $\lambda[T] = 3,\ 6$. However, $\lambda[S+T] = \frac{1}{2}(11 \pm \sqrt{37})$.

4.5. If $Au = \lambda u$, then

$$u = (A^{-1}A)u = A^{-1}(Au) = A^{-1}\lambda u \iff A^{-1}u = \frac{1}{\lambda}u.$$

4.6. Yes this is true. For a diagonolizable matrix $tA = S^{-1}\operatorname{diag}(t\lambda_1, \ldots, t\lambda_n)S$ the matrix exponential is defined as

$$e^{tA} = S^{-1}\operatorname{diag}(e^{t\lambda_1}, \ldots, e^{t\lambda_n})S.$$

4.7. **a)** The eigenvalues are $\lambda[T_{\Delta x}] = \frac{2}{\Delta x^2}(\mathrm{Re}(e^{i\pi k/(N+1)} - 1))$, which are the "$x$-values" of uniformly spaced points on a semicircle with radius $2/\Delta x^2$ centered at $x = -2/\Delta x^2$. The figure below demonstrates this.

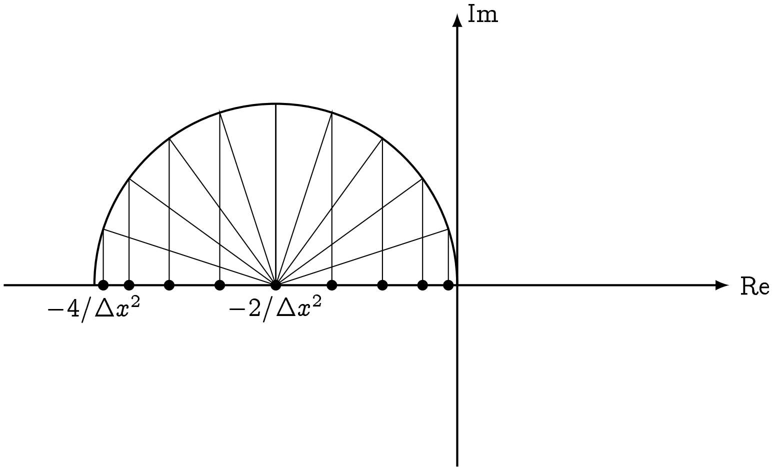

Figure 1: The eigenvalues $\lambda[T_{\Delta x}] = \frac{2}{\Delta x^2}(\cos(\pi k/(N+1)) - 1)$ are represented by •.

b) There is no good way of visualizing the eigenvalues of the inverse matrix. Let $\Delta x = 1/(N+1)$ and rewrite

$$2(\cos(\pi k/(N+1)) - 1) = -4\sin^2\left(\frac{k\pi}{2(N+1)}\right)$$

The eigenvalues for $T_{\Delta x}$ become (use $2\sin^2\theta = 1 - \cos(2\theta)$)

$$\frac{2}{\Delta x^2}(\cos(\pi k/(N+1)) - 1) = -4(N+1)^2\sin^2\left(\frac{k\pi}{2(N+1)}\right). \tag{1}$$

Another useful way (which we will use in later exercises) is to write the eigenvalues as

$$-\frac{4}{\Delta x^2} \cdot \sin^2(k\pi\Delta x/2). \tag{2}$$

For the inverse matrix we therefore get

$$\lambda\left[T_{\Delta x}^{-1}\right] = 1/\lambda[T_{\Delta x}] = -\frac{1}{4(N+1)^2\sin^2(k\pi/[2(N+1)])}$$

As we'll see in the next subproblem these eigenvalues will all cluster in a small region when $N \to \infty$ (we already see in the expression above that the values are rather small).

c) As $N \to \infty$ the eigenvalues of the "regular" matrix $T_{\Delta x}$ given by (1) approach

$$\lim_{N\to\infty} -4(N+1)^2\sin^2\left(\frac{k\pi}{2(N+1)}\right) = \lim_{N\to\infty} -4(N+1)^2\frac{k^2\pi^2}{4(N+1)^2} = -k^2\pi^2.$$

For the inverse we therefore get a small interval in which all eigenvalues cluster.

$$\lambda\left[T_{\Delta x}^{-1}\right] \approx -\frac{1}{k^2\pi^2} \implies \lambda\left[T_{\Delta x}^{-1}\right] \in (-1/\pi^2, 0).$$

4.8. **a)** A property of the logarithmic norm is that for any matrix A and $t \geq 0$ it holds that $||e^{tA}||_2 \leq e^{t\mu_2[A]}$. Applying this to $T_{\Delta x}$, which is symmetric, we get $\mu_2[T_{\Delta x}] = t\lambda_1$ (the largest eigenvalue given by (1)) and thus

$$||e^{tT_{\Delta x}}||_2 \leq e^{t\lambda_1} = e^{-4(N+1)^2 \sin^2(\pi/[2(N+1)])t} \approx e^{-\pi^2 t}$$

In fact, for symmetric, positive definite (all eigenvalues are positive) matrices A it is the case that $||A||_2 = \mu_2[A]$. Thus the bound given above is sharp for all $t \geq 0$.

b) As in the previous exercise, drawing won't give any useful information. As the next subproblem suggests, the eigenvalues will be approximately $e^{-tk^2\pi^2} \in (0, e^{-t\pi^2})$, which is a small interval (depending on t) on the positive real axis.

c) From the previous exercise we know the eigenvalues of $T_{\Delta x}$ are approximately $-k^2\pi^2$. Therefore the eigenvalues of $e^{tT_{\Delta x}}$ will approximately be $e^{-tk^2\pi^2}$.

d) If Δx is fixed then so is N. Therefore

$$\lim_{t\to\infty} \lambda[e^{tT_{\Delta x}}] = \lim_{t\to\infty} e^{-4t(N+1)^2 \sin^2(k\pi/[2(N+1)])} = 0.$$

e) Here we have the matrix $-T_{\Delta x}$ for which the largest eigenvalue is $-\lambda_N$, where λ_N is given by (1). The bound becomes

$$||e^{-tT_{\Delta x}}||_2 \leq e^{-t\lambda_N} = e^{4(N+1)^2 \sin^2(N\pi/[2(N+1)])t} \approx e^{N^2\pi^2 t},$$

which is unbounded as $N \to \infty$. Just like in the previous problem, the bound is always sharp since $e^{-tT_{\Delta x}}$ is symmetric and positive definite. As $\Delta x \to 0$ we therefore get $||e^{-tT_{\Delta x}}||_2 \to \infty$.

f) Solving $\dot{u} = T_{\Delta x} u$ with explicit Euler method gives the following time discretization

$$u_{t+1} = u_t + \Delta t \cdot T_{\Delta x} u_t = (I + \Delta t \cdot T_{\Delta x})u_t,$$

which is stable iff $|\lambda[I + \Delta t \cdot T_{\Delta x}]| \leq 1$. For these eigenvalues, it is the case that

$$\lambda[I + \Delta t \cdot T_{\Delta x}] = 1 + \Delta t \cdot \lambda[T_{\Delta x}],$$

and $\lambda[T_{\Delta x}] < 0$ (see (1)). The above expression's absolute value is therefore only less than 1 if

$$-\Delta t \cdot \lambda[T_{\Delta x}] - 1 \leq 1 \iff \Delta t \leq -\frac{2}{\lambda[T_{\Delta x}]} \leq \frac{2}{|\lambda_N|},$$

where λ_N is the most negative ("smallest") eigenvalue, i.e.

$$\lambda_N = -4(N+1)^2 \sin^2\left(\frac{N\pi}{2(N+1)}\right) \approx -4(N+1)^2 = -4/\Delta x^2$$

for large N. The bound is therefore approximately

$$\Delta t \leq \frac{2}{4/\Delta x^2} = \frac{\Delta x^2}{2} \iff \frac{\Delta t}{\Delta x^2} \leq \frac{1}{2},$$

which is the so-called CFL-condition.

g) For the implicit Euler method, we instead use

$$u_{t+1} = u_t + \Delta t \cdot T_{\Delta x} u_{t+1} \iff u_{t+1} = (I - \Delta t \cdot T_{\Delta x})^{-1} u_t.$$

Calculating the eigenvalues of $(I - \Delta t \cdot T_{\Delta x})^{-1}$ yields

$$\lambda\left[(I - \Delta t \cdot T_{\Delta x})^{-1}\right] = \frac{1}{1 - \Delta t \cdot \lambda[T_{\Delta x}]} > 0$$

since $\lambda[T_{\Delta x}] < 0$. The stability condition is therefore

$$\frac{1}{1 - \Delta t \cdot \lambda[T_{\Delta x}]} \leq 1 \iff \Delta t \geq 0.$$

In other words, the implicit method is always stable no matter the choice of Δt, which is its advantage.

h) The implicit method, since the CFL condition for the explicit method will require a small Δt if Δx is small. More precisely, if we half the size of Δx we will have to take a four times as small Δt.

4.9. The given expression is derived in the following way:

$$||u||_2^2 = \int_0^1 [u(t,x)]^2 \,\mathrm{d}x \iff$$

$$\frac{\partial ||u||_2^2}{\partial t} = \int_0^1 \frac{\partial}{\partial t}[u(t,x)]^2 \,\mathrm{d}x = \int_0^1 2 \cdot u(t,x) \cdot u_t(t,x) \,\mathrm{d}x = 2\langle u, u_t \rangle \iff \frac{1}{2}\frac{\partial ||u||_2^2}{\partial t} = \langle u, u_t \rangle.$$

Next, use the fact that $u_t = u_{xx}$ and do integration by parts.

$$\frac{1}{2}\frac{\partial ||u||_2^2}{\partial t} = \langle u, u_t \rangle = \langle u, u_{xx} \rangle = \underbrace{[u \cdot u_x]_{x=0}^{x=1}}_{0} - \int_0^1 [u_x]^2 \,\mathrm{d}x = -||u_x||_2^2 \iff \frac{1}{2}\frac{\partial ||u||_2^2}{\partial t} + ||u_x||_2^2 = 0.$$

Now, Sobolev's lemma tells us that $||u_x||_2^2 \geq \pi^2 ||u||_2^2$. Use this fact in the above result together with the substitution $y(t) = ||u(t,\cdot)||_2^2$ to get

$$\frac{1}{2}\frac{\mathrm{d}y}{\mathrm{d}t} + \pi^2 y \leq 0,$$

which is a differential inequality. If we had equality, the solution would be

$$y(t) = e^{-2\pi^2 t} \cdot y(0).$$

Since the derivative in our inequality is smaller than the above solution, the solution to the inequality decays faster than the above solution. Therefore

$$y(t) \leq e^{-2\pi^2 t} \cdot y(0) \iff ||u(t,\cdot)||_2^2 \leq e^{-2\pi^2 t} \cdot ||u(0,\cdot)||_2^2.$$

4.10. This matrix represents the difference relation

$$y_{n+1} - y_{n-1} = \lambda y_n, \quad y_0 = y_{N+1} = 0.$$

Use Viétes theorem on the characteristic equation:

$$r^2 - 1 = \lambda r \iff r^2 - \lambda r - 1 \iff \begin{cases} r_1 + r_2 = \lambda, \\ r_1 r_2 = -1. \end{cases}$$

The second equation yields $r_1 = -r_2^{-1} = r$. This implies that $r_1 = i\omega, r_2 = i/\omega$ and the general solution is therefore

$$y_n = A(i\omega)^n + B(i/\omega)^n = i^n \left(A\omega^n + B\omega^{-n}\right).$$

The boundary conditions yield

$$\begin{aligned} y_0 = A + B = 0 \qquad\qquad\qquad\qquad\qquad\quad &\implies y_n = Ai^n\left(\omega^n + \omega^{-1}\right), \\ y_{N+1} = Ai^{N+1}\left(\omega^{N+1} - \omega^{-(N+1)}\right) = 0 &\implies \omega^{N+1} - \omega^{-(N+1)} = 0. \end{aligned}$$

The last equation can be used to solve for ω.

$$\begin{aligned}\omega^{N+1}-\omega^{-(N+1)}&=0\\ \omega^{2(N+1)}&=1=e^{i2\pi k}\\ \omega_k&=e^{i\pi k/(N+1)}.\end{aligned}$$

Now we obtain λ_k from the first equation in Viéta's theorem.

$$\lambda_k=r_{1,k}+r_{2,k}=i\left(\omega_k+\omega_k^{-1}\right)=i\left(e^{i\pi k/(N+1)}+e^{-i\pi k/(N+1)}\right)=2i\cos(\pi k/(N+1))$$

4.11. **a)** Use the previous exercise and the fact that $\lambda[cA]=c\lambda[A]$ for any matrix A and constant c to get

$$\lambda[S_{\Delta x}]=\lambda[S/(2\Delta x)]=\frac{1}{2\Delta x}\lambda[S]=\frac{i}{\Delta x}\cos(\pi k/(N+1))=\frac{i}{\Delta x}\cos(\Delta x\pi k).$$

b) Since $S_{\Delta x}$ is skew-symmetric, i.e. $S_{\Delta x}=-S_{\Delta x}^T$, it is also *normal*, i.e. $S_{\Delta x}^T S_{\Delta x}=S_{\Delta x}S_{\Delta x}^T$. For normal matrices A it is the case that $||A||_2=\max_k|\lambda_k|$. In our case,

$$||S_{\Delta x}||_2=|\lambda_1|=\left|\frac{i}{\Delta x}\cos(\pi/(N+1))\right|=\frac{1}{\Delta x}\cos(\pi/(N+1)).$$

c) For normal matrices A it also holds that $\mu_2[A]=\max_k\{\mathrm{Re}(\lambda_k)\}$, which in our case means $\mu_2[S_{\Delta x}]=0$ since all eigenvalues are purely imaginary.

4.12. **a)** Use the fact that $\mu_2[\frac{\mathrm{d}^2}{\mathrm{d}x^2}+\frac{\mathrm{d}}{\mathrm{d}x}+1]\le\mu_2[\frac{\mathrm{d}^2}{\mathrm{d}x^2}]+\mu_2[\frac{\mathrm{d}}{\mathrm{d}x}]+\mu[1]$. For $\mu_2[\frac{\mathrm{d}^2}{\mathrm{d}x^2}]$ we have

$$\frac{\langle u'',u\rangle}{\langle u,u\rangle}=\frac{-\langle u',u'\rangle}{\langle u,u\rangle}=-\frac{\langle u',u'\rangle}{\langle u,u\rangle}=-\frac{||u'||_2^2}{||u||_2^2}\le-\frac{\pi^2||u||_2^2}{||u||_2^2}=-\pi^2.$$

To get the inequality we used Sobolev's lemma: $||u'||_2\ge\pi||u||_2$. Thus, $\mu_2[\frac{\mathrm{d}^2}{\mathrm{d}x^2}]=-\pi^2$. For $\mu_2[\frac{\mathrm{d}}{\mathrm{d}x}]$ we get

$$\frac{\langle u',u\rangle}{\langle u,u\rangle}=\frac{-\langle u,u'\rangle}{\langle u,u\rangle},$$

but since u is real-valued $\langle u',u\rangle=\langle u,u'\rangle$ and we have

$$\langle u',u\rangle=-\langle u,u'\rangle\implies\langle u',u\rangle=0.$$

Therefore, $\mu_2[\frac{\mathrm{d}}{\mathrm{d}x}]=0$. Calculating $\mu_2[1]=1$ is trivial and in total we get

$$\mu_2\left[\frac{\mathrm{d}^2}{\mathrm{d}x^2}+\frac{\mathrm{d}}{\mathrm{d}x}+1\right]\le 1-\pi^2.$$

To verify uniqueness, let $\mathcal{A}=\frac{\mathrm{d}^2}{\mathrm{d}x^2}+\frac{\mathrm{d}}{\mathrm{d}x}+1$ and $v\ne u$ another solution $\mathcal{A}v=f$ with $v(0)=v(1)=0$. Then the difference $u-v$ is a solution to the homogeneous equation, since

$$\begin{aligned}\mathcal{A}u=f,\quad\mathcal{A}v=f&\implies\mathcal{A}(u-v)=\mathcal{A}u-\mathcal{A}v=f-f=0,\\ u(0)-v(0)=0-0=u(1)-v(1)&\implies\text{Boundary conditions satisfied.}\end{aligned}$$

Since $\mu_2[\mathcal{A}]<0$, an inverse $\mathcal{A}^{-1}$ exists and therefore

$$\mathcal{A}(u-v)=0\iff u-v=\mathcal{A}^{-1}\cdot 0=0\iff u=v,$$

which is contradiction. The solution is therefore unique.

b) Let $h = 1/(N+1)$ and use the FDM

$$\frac{u_{n+1} - 2u_n + u_{n-1}}{h^2} + \frac{u_{n+1} - u_{n-1}}{2h} + u_n = f(x_n),$$

where $x_n = n \cdot h,\ n = 0, \dots, N+1$ and $u_0 = u_{N+1} = 0$. The corresponding matrix V_h is

$$V_h = \frac{1}{h^2} \cdot \operatorname{tridiag} \begin{pmatrix} 1 & -2 & 1 \end{pmatrix} + \frac{1}{2h} \cdot \operatorname{tridiag} \begin{pmatrix} -1 & 0 & 1 \end{pmatrix} + I$$

and the discretized problem becomes

$$V_h \boldsymbol{u} = \boldsymbol{f}, \qquad \boldsymbol{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{pmatrix}, \quad \boldsymbol{f} = \begin{pmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_N) \end{pmatrix}.$$

Since V_h is invertable, the system has a unique solution.

c) Here, $||\cdot||_{\Delta x}$ is the RMS-norm. We use the notation $||\cdot||_h$ instead and use the definition

$$||V_h||_h = \sqrt{h} \cdot ||V_h||_2.$$

We will use the logarithmic norm to give a bound to $||V_h||_2$ and therefore $||V_h||_h$. Unfortunately, we cannot directly compute $\mu_2[V_h]$, since the eigenvalues of V_h are unknown. However, we can apply the triangle inequality to get around this.

$$\mu_2[V_h] \leq \mu_2[S_h] + \mu_2[T_h] + \mu_2[I] = \max \operatorname{Re}(\lambda[S_h]) + \max \operatorname{Re}(\lambda[T_h]) + 1 = -\frac{4}{h^2} \cdot \sin^2(\pi h/2) + 1,$$

where $\lambda[S_h]$ and $\lambda[T_h]$ were calculated **4.11a)** and (2) respectively. Now, taking the operator 2-norm of the equation $\boldsymbol{u} = V_h^{-1}\boldsymbol{f}$ results in

$$||\boldsymbol{u}||_2 = ||V_h^{-1}\boldsymbol{f}||_2 \leq ||V_h^{-1}||_2 \cdot ||\boldsymbol{f}||_2 \leq -\frac{||\boldsymbol{f}||_2}{\mu_2[V_h]} \leq \frac{||\boldsymbol{f}||_2}{4\sin^2(\pi h/2)/h^2 - 1}.$$

In the last inequality we traded the minus sign for division:

$$\frac{||\boldsymbol{f}||_2}{\mu_2[V_h]} \geq \frac{||\boldsymbol{f}||_2}{1 - 4\sin^2(\pi h/2)/h^2} \iff -\frac{||\boldsymbol{f}||_2}{\mu_2[V_h]} \leq -\frac{||\boldsymbol{f}||_2}{1 - 4\sin^2(\pi h/2)/h^2}.$$

Multiplying both sides by $\sqrt{h}$ yields

$$||\boldsymbol{u}||_h \leq \frac{||\boldsymbol{f}||_h}{4\sin^2(\pi h/2)/h^2 - 1}.$$

4.13. **a)** For the logarithmic 2-norm of $\mathcal{A} = \frac{\mathrm{d}^2}{\mathrm{d}x^2} + \omega^2$ we get

$$\mu_2[\mathcal{A}] \leq \mu_2\left[\frac{\mathrm{d}^2}{\mathrm{d}x^2}\right] + \mu_2\left[\omega^2\right] = \omega^2 - \pi^2.$$

When $\mu_2[\mathcal{A}] < 0$ the inverse operator $\mathcal{A}^{-1}$ exists and the solution is therefore unique. This condition is guaranteed if $|\omega| < \pi$. The condition is therefore $-\pi < \omega < \pi$.

b) The problem will have multiple solutions. To see this, examine the homogeneous equation

$$\frac{\mathrm{d}^2 y_h}{\mathrm{d}x^2} + \pi^2 y_h = 0, \quad y_h(0) = y_h(1) = 0.$$

This equation has the general solution $y_h(x) = a\cos(\pi x) + b\sin(\pi x)$. The first boundary condition results in

$$y_h(0) = a = 0 \implies y_h(x) = b\sin(\pi x).$$

However, the second boundary condition holds for any b since $\sin(\pi) = 0$. The homogeneous equation therefore has infinitely many solutions $y_h(x) = b\sin(\pi x)$. If we have any particular equation $\mathcal{A}y_p = g(x)$ with a solution y_p we can always add the homogeneous solution y_h to it and get a new solution. Therefore, the equation $\mathcal{A}y = g(x)$ with $\omega = \pi$ will have infinitely many solutions for any g.

5 Partial Differential Equations I

5.1. A strategy to show that a problem depends continuously on the initial condition is to find a upper bound for the norm of the solution which depends on the norm of the initial condition as well as time. In our case, we can use the result from exercise **4.9**:

$$||u(t,\cdot)||_2^2 \leq e^{-2\pi^2 t}||u(0,\cdot)||_2^2 = e^{-2\pi^2 t}||g||_2^2.$$

Now, suppose we want to solve the same problem, but with a slightly perturbed initial condition $u(0,x) = h(x)$ such that

$$||g - h||_2^2 < \varepsilon$$

for some (arbitrarily) small ε. Let the solution to this problem be $v(t,x)$ (we won't concern ourselves with the existence of this solution). Then, the difference $w(t,x) = u(t,x) - v(t,x)$ satisfies

$$w_t = w_{xx}, \quad w(t,0) = w(t,1) = 0, \quad w(0,x) = g(x) - h(x)$$

because the PDE is linear. Now, apply the result from problem **4.9** to get

$$||w(t,\cdot)||_2^2 = ||u(t,\cdot) - v(t,\cdot)||_2^2 \leq e^{-2\pi^2 t}||g - h||_2^2 < \varepsilon e^{-2\pi^2 t},$$

which goes to zero as $\varepsilon \to 0$ for any fixed t. This shows the solution's continuous dependence on the initial condition, since for any given $t, \varepsilon > 0$, we can pick $\delta = \varepsilon e^{-2\pi^2 t}$ such that

$$||g - h||_2^2 < \varepsilon \implies ||u(t,\cdot) - v(t,\cdot)||_2^2 < \delta.$$

5.2. We know the eigenfunctions and -values of $-\frac{\partial^2}{\partial x^2}$ over $[0,1]$ with homogeneous Dirichlet-conditions are

$$\varphi_k(x) = \sin(k\pi x), \quad \lambda_k = k^2\pi^2.$$

We do the ansatz $u(t,x) = u_k(t)\varphi_k(x)$ and get

$$\frac{\partial u}{\partial t} = u_k'(t)\varphi_k(x) = -\frac{\partial^2 u}{\partial x^2} = u_k(t)\lambda_k\varphi_k(x) \iff u_k'(t) = \lambda u_k(t) \iff u_k(t) = c_k e^{\lambda_k t}$$

Thus, the functions $u_n(t,x) = c_n e^{n^2\pi^2 t}\sin(n\pi x)$ are all solutions to the problem. If we pick $c_n = n^r$ with $r < 0$ we get (note that $||\sin(n\pi x)||_2 = \sqrt{2}$)

$$||u_n(0,x)||_2 = \sqrt{2}n^r \to 0 \quad \text{as } n \to \infty.$$

This means that the initial condition gets arbitrarily small for large n. For the solution $u_n(t,x)$ however, we get

$$||u_n(t,x)||_2 = \sqrt{2}n^r e^{n^2\pi^2 t} \to \infty \quad \text{as } n \to \infty$$

for any fixed t because exponential functions grow quicker than power functions. In other words, as we make the initial condition smaller and smaller, the solution blows up to infinity, which proves that the problem is ill-posed by counter-example.

5.3. The problem has the solution

$$u(t) = e^{tT_{\Delta x}} u(0).$$

With perturbed initial condition $v(0) = u(0) + \varepsilon$, the solution v is

$$v(t) = e^{tT_{\Delta x}} v(0) = e^{tT_{\Delta x}} u(0) + e^{tT_{\Delta x}} \varepsilon.$$

The solution is therefore perturbed by $e^{tT_{\Delta x}}\varepsilon$.

5.4. The perturbation is now instead $e^{-tT_{\Delta x}}\varepsilon$. Since the largest eigenvalue of $e^{-tT_{\Delta x}}$ is approximately $e^{N^2\pi^2 t}$, which for fixed t grows arbitrarily large as $N \to \infty$, the perturbation will also grow arbitrarily large for any given ε. Thus, we conclude that the problem is ill-posed.

5.5. This can easily be proved using powerful results from matrix theory. Let J be the Jordan normal form of A, i.e. $A = S^{-1}JS$ for some invertable S. The the matrix function $Q(A)$ is defined as (it can also be algebraically derived)

$$Q(A) = S^{-1}Q(J)S \implies \lambda[Q(A)] = \lambda[Q(J)].$$

Thus, we only have to prove this for a Jordan matrix J. For any matrix it is the case that

$$\lambda[(\beta I + \gamma J)^{-1}] = 1/(\beta + \gamma\lambda[J]),$$
$$\lambda[I + \alpha J] = 1 + \alpha\lambda[J]).$$

The last piece of the puzzle is to show that

$$\lambda[(\beta I + \gamma J)^{-1}(I + \alpha J)] = \lambda[(\beta I + \gamma J)^{-1}] \cdot \lambda[I + \alpha J].$$

But for a Jordan matrix this is trivial, because the matrices in question are upper triangular. For upper triangular matrices A, B the eigenvalues are the diagonal elements, i.e. $\lambda[A] = a_{ii}, \lambda[B] = b_{ii}$, and the product $C = AB$ is also upper triangular with diagonal elements $c_{ii} = a_{ii}b_{ii}$. Combining these facts shows the identity stated above and therefore

$$\lambda[Q(J)] = \lambda[(\beta I + \gamma J)^{-1}(I + \alpha J)] = \lambda[(\beta I + \gamma J)^{-1}] \cdot \lambda[I + \alpha J] = \frac{1 + \alpha\lambda[J]}{\beta + \gamma\lambda[J]} = Q(\lambda[J]).$$

5.6. Using the previous exercise, we get

$$Q(w) = \frac{1 + w\Delta t/2}{1 - w\Delta t/2} \implies \lambda[B(\Delta t, \Delta x)] = Q(\lambda[T_{\Delta x}]) = \frac{1 + \lambda[T_{\Delta x}]\Delta t/2}{1 - \lambda[T_{\Delta x}]\Delta t/2}$$

with $\lambda[T_{\Delta x}] = -4(N+1)^2 \sin^2(k\pi/[2(N+1)])$, see (1). The condition for stability is

$$|\lambda[B(\Delta t, \Delta x)]| \le 1 \implies \frac{|1 + \lambda[T_{\Delta x}]\Delta t/2|}{|1 - \lambda[T_{\Delta x}]\Delta t/2|} \le 1,$$

which is the same inequality we got in exercise **1.13**. It is equivalent to the inequality

$$\text{Re}(\lambda[T_{\Delta x}]\Delta t/2) \le 0,$$

which holds for any $\Delta t \ge 0$, since $\lambda[T_{\Delta x}] < 0$. Thus, there is no restriction on Δt.

5.7. **a)** $\mu_2[R_{\Delta x}] = \mu_2[-R_{\Delta x}] = 0$, since the eigenvalues are purely imaginary.

b) The solution to $\dot{u} = R_{\Delta x} u$ with $u(0) = u_0$ is

$$u = e^{tR_{\Delta x}} u_0 \implies ||u||_2 = ||e^{tR_{\Delta x}} u_0||_2 \leq ||e^{tR_{\Delta x}}|| \cdot ||u_0||_2 \leq \underbrace{e^{t\mu[R_{\Delta x}]}}_{1} \cdot ||u_0||_2 = ||u_0||_2.$$

Using the same argument as in exercise **5.1**, we conclude that the problem is well-posed.

Full solution: Replace u by $u - v$ in the above result and let $||u_0 - v_0||_2 < \varepsilon$ for some small ε. Then

$$||u_0 - v_0||_2 < \varepsilon \implies ||u - v||_2 \leq ||u_0 - v_0|| < \delta$$

with $\delta = \varepsilon$.

In reverse time, the result is the same because $\mu_2[R_{\Delta x}] = \mu_2[-R_{\Delta x}]$. The problem is therefore well-posed in both forward and reverse time.

5.8. From the calculations done in problem **4.12a)** we know that

$$\mu_2\left[\frac{\mathrm{d}^2}{\mathrm{d}x^2} + K\frac{\mathrm{d}}{\mathrm{d}x} + 1\right] \leq 1 - \pi^2,$$

which is both negative and independent of K. Abbreviate $\mathcal{A} = \frac{\mathrm{d}^2}{\mathrm{d}x^2} + K\frac{\mathrm{d}}{\mathrm{d}x} + 1$. Since $\mu_2[\mathcal{A}] < 0$ there exists an inverse operator $\mathcal{A}^{-1}$ and the unique solution to the equation is therefore $y = \mathcal{A}^{-1}g$. Taking the norm of both sides of the equation and using the properties

$$\mu_2[\mathcal{A}] < 0 \implies ||\mathcal{A}^{-1}||_2 \leq -\frac{1}{\mu_2[\mathcal{A}]}, \quad ||\mathcal{A}u||_2 \leq ||\mathcal{A}||_2 \cdot ||u||_2$$

yields

$$||y||_2 = ||\mathcal{A}^{-1}g||_2 \leq ||\mathcal{A}^{-1}||_2 \cdot ||g||_2 \leq -\frac{1}{\mu_2[\mathcal{A}]} \cdot ||g||_2 = \frac{||g||_2}{\pi^2 - 1}.$$

As in previous problems, we conclude from the above that the problem is well-posed (independently of K).

5.9. We use the same strategy as in the last problem. This time, however, we have $\mu_2[\mathcal{A}] \leq K - \pi^2$, which is guaranteed to be negative only for $K < \pi^2$. If $\mu_2[\mathcal{A}]$ is non-negative, the solution is not unique anymore and the problem is then ill-posed. Hence, the value of K matters for this problem.

5.10. This is pretty much a combination of problems **4.9** and **5.1**. A full solution would look something like this: Start by deriving the squared 2-norm of u.

$$\frac{\partial}{\partial t}||u||_2^2 = \frac{\partial}{\partial t}\int_0^1 u^2\,\mathrm{d}x = \int_0^1 \frac{\partial}{\partial t}u^2\,\mathrm{d}x = \int_0^1 2uu_t\,\mathrm{d}x = 2\langle u, u_t\rangle,$$

since the scalar product is defined as $\langle u, v\rangle = \int_0^1 u\overline{v}\,\mathrm{d}x$. Now, from the given equation we know

$$u_t = u_x + \frac{u_{xx}}{\mathsf{Pe}}.$$

Inserting this into the scalar product above yields

$$2\langle u, u_t\rangle = 2\langle u, u_x + \frac{u_{xx}}{\mathsf{Pe}}\rangle = 2\underbrace{\langle u, u_x\rangle}_{=0} + \frac{2}{\mathsf{Pe}} \cdot \langle u, u_{xx}\rangle = \frac{2}{\mathsf{Pe}} \cdot \langle u, u_{xx}\rangle = -\frac{2}{\mathsf{Pe}} \cdot ||u_x||_2^2,$$

since $\langle u, u_{xx}\rangle = -\langle u_x, u_x\rangle = -||u_x||_2^2$. After combining the above results and using Sobolev's lemma we get (similar to problem **4.9**)

$$\frac{\mathrm{d}||u||_2^2}{\mathrm{d}t} + \frac{2}{\mathrm{Pe}} \cdot ||u_x||_2^2 = 0 \iff \frac{\mathrm{d}||u||_2^2}{\mathrm{d}t} + \frac{2\pi^2}{\mathrm{Pe}} \cdot ||u||_2^2 \leq 0.$$

The solution to the differential inequality is

$$||u(t,x)||_2^2 \leq e^{-2\pi^2 t/\mathrm{Pe}} \cdot ||u(0,x)||_2^2$$

and by the same argument as in problem **5.1** we can conclude that the problem is well-posed.

5.11. Reverse time means that our equation instead becomes

$$u_t = -u_x - \frac{u_{xx}}{\mathrm{Pe}},$$

i.e. the right hand side switches sign. If we run through the procedure from the previous problem with this equation again we see that replacing u_x with $-u_x$ doesn't change anything, since $\langle u, u_x\rangle = -\langle u, u_x\rangle = 0$. The culprit is the $\langle u, u_{xx}\rangle$ term which now is negative and applying Sobolev's lemma will now give us a lower bound instead of an upper bound:

$$\frac{\mathrm{d}||u||_2^2}{\mathrm{d}t} - \frac{2}{\mathrm{Pe}} \cdot ||u_x||_2^2 = 0 \iff \frac{\mathrm{d}||u||_2^2}{\mathrm{d}t} - \frac{2\pi^2}{\mathrm{Pe}} \cdot ||u||_2^2 \geq 0.$$

It is here the argument from the previous exercise breaks down. This does, of course, not prove that the problem is ill-posed. To do this, we will have to do something similar to problem **5.2**. I will not provide the rest of the solution, as this material is not part of the course any more, and because a solution to a similar problem already exists as mentioned.

5.12. Insert $u(x,y) = X(x)Y(y)$ into $\Delta u = \lambda u$ and get

$$X''Y + XY'' = \lambda XY \iff \frac{X''}{X} + \frac{Y''}{Y} = \lambda.$$

The first term only depends on x, while the second only depends on y and the right hand side is constant. From that, we conclude that each term on the left hand side must be constant. Granted we know that λ is negative, let $X''/X = -\alpha, Y''/Y = -\beta$ for some positive α, β. From that we get the three equations

$$\begin{cases} X'' + \alpha X = 0, \quad X(0) = X(1) = 0, \\ Y'' + \beta Y = 0, \quad Y(0) = Y(1) = 0, \\ \lambda = -\alpha - \beta. \end{cases}$$

The first two equations are exactly the same and are simply the eigenvalue problem for $\frac{\mathrm{d}^2}{\mathrm{d}x^2}$ over $[0,1]$ with homogeneous Dirichlet conditions. They have the solutions

$$\begin{cases} X_i(x) = a_i \sin(i\pi x), \quad \alpha_i = (i\pi)^2 = -\kappa_i, \\ Y_j(y) = b_j \sin(j\pi y), \quad \beta_j = (j\pi)^2 = -\kappa_j \end{cases}$$

with $i, j > 0$. From that we conclude that the eigenvalues for the 2-dimensional Laplacian are

$$\lambda_{i,j} = -\alpha - \beta = \kappa_i + \kappa_j = -\pi^2(i^2 + j^2).$$

The largest eigenvalue is $\lambda_{1,1} = -2\pi^2$.

5.13. Let $h = \Delta x = \Delta y$ and use the shorthand $T(X_i) = h^{-2}(X_{i-1} - 2X_i + X_{i+1})$. Then with $u_{i,j} = X_iY_j$ we get

$$\begin{cases} \frac{u_{i-1,j}-2u_{i,j}+u_{i+1,j}}{h^2} = T(X_i)Y_j, \\ \frac{u_{i,j-1}-2u_{i,j}+u_{i,j+1}}{h^2} = X_iT(Y_j). \end{cases} \iff T(X_i)Y_j + T(X_i)Y_j = \lambda X_iY_j \iff \frac{T(X_i)}{X_i} + \frac{T(Y_j)}{Y_j} = \lambda.$$

Using the same argument as in the previous exercise, we generate the equations

$$\begin{cases} T(X_i) + \alpha X_i = 0, \quad X_0 = X_{N+1} = 0, \\ T(Y_j) + \beta Y_j = 0, \quad Y_0 = Y_{M+1} = 0, \\ \lambda = -\alpha - \beta. \end{cases}$$

The first two equations can be put into matrix form and will result in eigenvalue problems for the well-studied matrix

$$T_h = h^{-2}\,\mathrm{tridiag}\begin{pmatrix}1 & -2 & 1\end{pmatrix}.$$

From this matrix we get the eigenvalues

$$\alpha_i = \frac{4}{h^2}\sin^2(i\pi h/2), \quad \beta_j = \frac{4}{h^2}\sin^2(j\pi h/2).$$

The eigenvalues $\lambda_{i,j}$ for the original problem therefore become

$$\lambda_{i,j} = -\alpha_i - \beta_j = -\frac{4}{h^2}(\sin^2(i\pi h/2) + \sin^2(j\pi h/2)) \approx -\pi^2(i^2 + j^2)$$

as expected.

5.14. Since $L_{\Delta x}$ is symmetric, we use the fact

$$\mu_2[L_{\Delta x}] = \max_k\{\lambda_k\} = -\frac{8}{h^2}\sin^2(\pi h/2).$$

Here, we used the largest eigenvalue $\lambda_{1,1}$ from the previous exercise.

6 Partial Differential Equations II

6.1. Use the same strategy as in problem **5.1**. Since this is the advection equation, we have to assume periodic boundary conditions, i.e. $u(a) = u(b)$ (we solve the equation over some interval $[a, b]$). Derive $||u||_2^2$ w.r.t. time and use the equation $u_t = u_x$ to get

$$\frac{\partial ||u||_2^2}{\partial t} = \int_a^b \frac{\partial}{\partial t}u^2\,\mathrm{d}x = 2\int_a^b u_tu\,\mathrm{d}x = 2\int_a^b u_xu\,\mathrm{d}x = 2\underbrace{[u^2]_a^b}_{=0} - 2\int_a^b uu_x\,\mathrm{d}x.$$

As usual, we see that $\int_a^b u_xu\,\mathrm{d}x = -\int_a^b u_xu\,\mathrm{d}x = 0$. Thus,

$$\frac{\partial ||u||_2^2}{\partial t} = 0 \implies ||u(t,\cdot)||_2 = ||u(0,\cdot)||_2$$

for any $t > 0$. Put into words, the norm of the solution is preserved over time. By the same argument as in **5.1** we conclude that the problem is well posed. The same is true in reverse time, since swapping u_x by $-u_x$ doesn't change the fact that $\int_a^b u_xu\,\mathrm{d}x = -\int_a^b u_xu\,\mathrm{d}x = 0$ and the result is therefore unchanged.

Full solution: Let $v_t = v_x$ for $t > 0, x \in [a, b]$ with $v(0, x) = h(x)$ and periodic boundary conditions such that

$$||g - h||_2^2 < \varepsilon,$$

where $g(x) = u(0, x)$ for some $\varepsilon > 0$. Then $w(t, x) = u(t, x) - v(t, x)$ is also a solution, since

$$w_t = \frac{\partial}{\partial t}(u - v) = u_t - v_t = u_x - v_x = w_x.$$

The initial condition is $w(0, x) = u(0, x) - v(0, x) = g(x) - h(x)$. From the above result we know that the norm of the solution is the same as the norm of the initial condition, i.e.

$$||w(t, \cdot)||_2 = ||w(0, \cdot)||_2 \iff ||u(t, \cdot) - v(t, \cdot)||_2 = ||g - h||_2 < \delta = \varepsilon.$$

Thus, we have shown that given a small change $\varepsilon > 0$ to the initial condition we can always find a $\delta > 0$ such that the solution is perturbed by something smaller than δ. In other words, the solution depends continuously on the initial data and the problem is well-posed.

6.2. Remember, in order for the recurrence relation $u_{n+1} = Au_n$ to have a bounded (stable) solution, it must be the case that $\rho(A) \leq 1$, where $\rho(A)$ is the spectral radius of A. In other words, *all eigenvalues must have (squared) magnitude less than or equal to one.* Also, from exercise **4.11a)** we know that the eigenvalues of $S_{\Delta x}$ are

$$\lambda[S_{\Delta x}] = \frac{i}{\Delta x}\cos(\Delta x \pi k).$$

a) In this case $A = I + \Delta t \cdot S_{\Delta x}$. The eigenvalues are

$$\lambda[A] = 1 + \Delta t \cdot \lambda[S_{\Delta x}] = 1 + i \cdot \frac{\Delta t}{\Delta x}\cos(\Delta x \pi k).$$

We already see that the method is unstable. If you aren't convinced, here's why:

$$|\lambda[A]|^2 = 1 + \frac{\Delta t^2}{\Delta x^2}\cos^2(\Delta x \pi k) > 1.$$

b) From the above it's not hard to see that this method will *always* be stable. It is because in this case, we have

$$\lambda[A] = \frac{1}{1 + \Delta t \cdot \lambda[S_{\Delta x}]} = \frac{1}{1 + i \cdot \frac{\Delta t}{\Delta x}\cos(\Delta x \pi k)}$$

and the magnitude of the denominator will be the same as for the matrix in the explicit method. In other words,

$$|\lambda[A]|^2 = \frac{1}{1 + \frac{\Delta t^2}{\Delta x^2}\cos^2(\Delta x \pi k)} \leq 1 \iff 1 + \frac{\Delta t^2}{\Delta x^2}\cos^2(\Delta x \pi k) \geq 1,$$

which is what we found from the explicit method.

c) Here, we use the results from problem **5.5** to compute the eigenvalues.

$$\lambda[A] = \frac{1 + \frac{\Delta t}{2} \cdot \lambda[S_{\Delta x}]}{1 - \frac{\Delta t}{2} \cdot \lambda[S_{\Delta x}]}.$$

From the previous methods we know that both the numerator and denominator have the same magnitude. Consequently, $|\lambda[A]| = 1$ for any $\Delta t, \Delta x > 0$ and the method is always stable.

6.3. Solving $u_t = u_x$ in reverse time is the same as solving $u_t = -u_x$ in forward time. Applying the methods from the previous exercise on this problem will therefore result in the same recurrence relations with the only difference being that $S_{\Delta x}$ is replaced by $-S_{\Delta x}$. It's quite obvious that the stability properties won't change because of this for any of the three methods above. For explicit Euler, the solution will explode in both forward and reverse time, while for the implicit Euler the solution will decay in both forward and reverse time. For the trapezoidal rule, however, we claim that the norm is always preserved, i.e.

$$||u_{n+1}||_2 = ||u_n||_2, \quad u_{n+1} = Au_n, \quad A = \left(I - \frac{\Delta t}{2} \cdot S_{\Delta x}\right)^{-1} \cdot \left(I + \frac{\Delta t}{2} \cdot S_{\Delta x}\right).$$

The reason is that A is unitary in this case and for unitary matrices A we know (from matrix theory) that $||AX||_2 = ||X||_2$ for any vector X. To prove this, recall that a matrix A is unitary iff $A^H A = I$. Next, note that

$$\left(I - \frac{\Delta t}{2} \cdot S_{\Delta x}\right)^H = I^H - \frac{\Delta t}{2} \cdot S_{\Delta x}^H = I + \frac{\Delta t}{2} \cdot S_{\Delta x} =: B,$$

$$\left(I + \frac{\Delta t}{2} \cdot S_{\Delta x}\right)^H = I^H + \frac{\Delta t}{2} \cdot S_{\Delta x}^H = I - \frac{\Delta t}{2} \cdot S_{\Delta x} =: C.$$

Moreover, the matrices B and C (and therefore also B^{-1} and C^{-1}) commute, since

$$BC = \left(I + \frac{\Delta t}{2} \cdot S_{\Delta x}\right)\left(I - \frac{\Delta t}{2} \cdot S_{\Delta x}\right) = I - \frac{\Delta t^2}{4} \cdot S_{\Delta x}^2 = \left(I - \frac{\Delta t}{2} \cdot S_{\Delta x}\right)\left(I + \frac{\Delta t}{2} \cdot S_{\Delta x}\right) = CB.$$

With these properties, together with the fact that $(A^{-1})^H = (A^H)^{-1}$, we are ready to prove the main result.

$$A^H A = \left(C^{-1}B\right)^H \cdot \left(C^{-1}B\right) = B^H (C^H)^{-1} C^{-1} B = CB^{-1}C^{-1}B = CC^{-1}B^{-1}B = I.$$

Now we are confident that $||u_{n+1}||_2 = ||u_n||_2$ when we use the trapezoidal rule, which is very good because the original equation behaves in the same way; $||u(t,\cdot)||_2 = ||u(0,\cdot)||_2$. This is therefore the most suitable method to use in this case.

6.4. **a)** The given expression is supposed to be applied to the equation $\dot{u} = S_{\Delta x} u$. With the linear test equation $\dot{u} = \lambda u$ instead, we simply replace $S_{\Delta x}$ with λ and get the recurrence relation

$$u_{n+1} = u_{n-1} + 2\Delta t \lambda u_n$$

which has the characteristic equation

$$r^2 = 1 + 2\Delta t \lambda r \iff r^2 - 2\Delta t \lambda r - 1 = 0.$$

b) This follows directly from Viéta's theorem, which states that the product of the roots of a quadratic polynomial is the constant term, which in this case is precisely -1.

c) Let r_1, r_2 be the roots of the characteristic polynomial. We know that $r_1 r_2 = -1$ and taking magnitude of both sides yields

$$|r_1| \cdot |r_2| = 1.$$

Without loss of generality, it's obvious that if $|r_1| < 1$ then

$$|r_2| = \frac{1}{|r_1|} > 1,$$

which violates the root condition and the method therefore becomes unstable.

d) As we've seen above, both roots must have unit magnitude in order to have stability. It is therefore the case that

$$r_1 = e^{i\theta}, \quad r_2 = -e^{-i\theta}.$$

Here, $\theta \neq \pi/2$, since in that case $r_1 = r_2 = i$ is a double root, which violates the root condition. Viéta's theorem also tells us that the sum of roots is the negative coefficient in front of the r-term, i.e.

$$r_1 + r_2 = e^{i\theta} - e^{-i\theta} = 2i\sin\theta = 2\Delta t\lambda \iff \Delta t\lambda = i\sin\theta$$

and $\sin\theta \in (-1, 1)$ since $\theta \neq \pi/2$. Thus,

$$\Delta t\lambda = i\omega, \quad \omega \in (-1, 1).$$

6.5. Diagonalize $S_{\Delta x}$ as $S_{\Delta x} = S^{-1}\Lambda S$. Then, by letting $v = Su$ we get

$$\dot{u} = S_{\Delta x}u \iff \dot{u} = S^{-1}\Lambda Su \iff S\dot{u} = \Lambda Su \iff \dot{v} = \Lambda v.$$

Denote $v_n = \begin{pmatrix} v_n^1 & v_n^2 & \cdots & v_n^N \end{pmatrix}^T$. Applying the Leap-frog on the above equation will result in

$$v_{n+1}^k = v_{n-1}^k + 2\Delta t \cdot \lambda_k[S_{\Delta x}]v_n^k$$

for component k.

a) From the previous exercise we know that the condition for stability is $\Delta t\lambda = i\omega$ for $\omega \in (-1, 1)$. In our case, $\lambda_k[S_{\Delta x}]$ is given by the result in problem **4.11a)** and therefore

$$\Delta t\lambda_k[S_{\Delta x}] = i \cdot \frac{\Delta t}{\Delta x}\cos(\Delta x\pi k) \implies \frac{\Delta t}{\Delta x} \cdot |\cos(\Delta x\pi k)| < 1$$

for stability. Consequently,

$$\frac{\Delta t}{\Delta x} < \frac{1}{|\cos(\Delta x\pi k)|} \geq \frac{1}{|\cos(\Delta x\pi)|} \approx 1$$

for small Δx. The CFL-condition is therefore $\Delta t \approx \Delta x$.

b) It's not clear what it means to "run the method in reverse". Assume that it means that we replace the given recurrence relation with

$$v_{n-1}^k = v_{n+1}^k + 2\Delta t \cdot \lambda_k[S_{\Delta x}]v_n^k.$$

Then the characteristic equation now becomes

$$1 = r^2 + 2\Delta t \cdot \lambda_k[S_{\Delta x}]r \iff r^2 + 2\Delta t \cdot \lambda_k[S_{\Delta x}]r - 1 = 0.$$

The only difference from the regular case is the change of sign in the r-term. Viéta's theorem tells us that

$$\begin{cases} r_1 + r_2 = -2\Delta t \cdot \lambda_k[S_{\Delta x}], \\ r_1 \cdot r_2 = -1. \end{cases}$$

The second equation is the same as before, hence $r_1 = e^{i\theta}, r_2 = -e^{-i\theta}$. But from the first equation we get

$$r_1 + r_2 = 2i\sin\theta = -2\Delta t \cdot \lambda_k[S_{\Delta x}] \iff \Delta t \cdot \lambda_k[S_{\Delta x}] = i\omega, \quad \omega \in (-1, 1),$$

i.e. the same result as before. In other words, the behaviour is the same when running in both forward and reverse time.

6.6. We get

$$\lim_{n\to\infty}\left(1+\frac{i\omega_k t/\Delta x}{n}\right)^n = e^{i\omega_k t/\Delta x}.$$

Since $\omega_k = \cos(\Delta x \pi k)$ (see problem **4.11a)**), $e^{i\omega_k t/\Delta x}$ will always have unit magnitude. However, as $\Delta x \to 0$ we get

$$\lim_{\Delta x\to 0} e^{i\omega_k t/\Delta x} = [\theta = \omega_k t/\Delta x] = \lim_{\theta\to\infty} e^{i\theta},$$

which is a limit that does not exist. In practice of course, we are not actually working with a complex exponential $e^{i\theta}$, only an approximation with some fixed n. The magnitude is then

$$\left|1+\frac{i\omega_k t/\Delta x}{n}\right|^n = \left(1+\frac{\omega_k^2 t^2}{n^2\Delta x^2}\right)^{n/2} \approx \left(1+\frac{\omega_k^2 t^2}{2n^2\Delta x^2}\right)^n \leq 1+\frac{\omega_k^2 t^2}{2n\Delta x^2}.$$

Here, we first used a Taylor expansion

$$\sqrt{1+x^2} \approx 1+\frac{x^2}{2}$$

for small x followed by only using the first two terms in the binomial expansion. We see that the upper bound is always greater than one, but in order to keep the magnitude error bound at a certain value

$$\varepsilon = \frac{\omega_k^2 t^2}{2n\Delta x^2}$$

one has to fix the value of

$$n\cdot\Delta x^2 \approx \frac{\omega_k^2 t^2}{2\varepsilon}.$$

6.7. Use the shorthands $u = u(t,x), u_t = u_t(t,x)$, etc. We use the Taylor expansion

$$u(t+\Delta t, x) = u + \Delta t u_t + \frac{\Delta t^2}{2}u_{tt} + \mathcal{O}(\Delta t^3)$$

and want to replace the temporal derivatives with spacial ones.

$$u_t + au_x = 0 \implies \begin{cases} u_{tt} + au_{xt} = 0, \\ u_{tx} + au_{xx} = 0, \end{cases} \implies u_{tt} + a\cdot(-au_{xx}) = 0 \implies \begin{cases} u_t = -au_x, \\ u_{tt} = a^2 u_{xx}. \end{cases}$$

Inserting this into the Taylor expansion yields

$$u(t+\Delta t, x) = u - a\Delta t u_x + \frac{a^2\Delta t^2}{2}u_{xx} + \mathcal{O}(\Delta t^3)$$

Then, use the second order approximations

$$u_x \approx \frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x},$$
$$u_{xx} \approx \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{\Delta x^2}$$

and insert into the expansion.

$$u_j^{n+1} = u_j^n - a\Delta t\cdot\frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} + \frac{a^2\Delta t^2}{2}\cdot\frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{\Delta x^2}.$$

Then, by letting $\mu = \Delta t/\Delta x$ we get

$$\begin{aligned} u_j^{n+1} &= u_j^n - \frac{a\mu}{2}\cdot(u_{j+1}^n - u_{j-1}^n) + \frac{a^2\mu^2}{2}\cdot(u_{j+1}^n - 2u_j^n + u_{j-1}^n) \\ &= \frac{a\mu}{2}(a\mu-1)u_{j+1}^n + (1-a^2\mu^2)u_j^n + \frac{a\mu}{2}(a\mu+1)u_{j-1}^n. \end{aligned}$$

6.8. With periodic boundary conditions we will always get circulant matrices. In our case we will get the matrices

$$\frac{u_{j+1}^n + u_{j-1}^n}{2} \implies \frac{1}{2}\cdot \operatorname{Circ}(0,1,\dots,1) = \frac{1}{2}\cdot\begin{pmatrix} 0 & 1 & 0 & \cdots & 0 & 1\\ 1 & 0 & 1 & \cdots & 0 & 0\\ 0 & 1 & 0 & \cdots & 0 & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & 0 & \cdots & 0 & 1\\ 1 & 0 & 0 & \cdots & 1 & 0\end{pmatrix},$$

$$a\Delta t\cdot\frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} \implies \frac{a\Delta t}{2\Delta x}\cdot \operatorname{Circ}(0,1,\dots,-1) = \frac{a\Delta t}{2\Delta x}\cdot\begin{pmatrix} 0 & 1 & 0 & \cdots & 0 & -1\\ -1 & 0 & 1 & \cdots & 0 & 0\\ 0 & -1 & 0 & \cdots & 0 & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & 0 & \cdots & 0 & 1\\ 1 & 0 & 0 & \cdots & -1 & 0\end{pmatrix}.$$

Adding these matrices will result in another circulant matrix

$$A = \operatorname{Circ}\left(0, \frac{1}{2} - \frac{a\Delta t}{2\Delta x}, \dots, \frac{1}{2} + \frac{a\Delta t}{2\Delta x}\right).$$

And for CFL $= 1$, we get

$$\frac{a\Delta t}{\Delta x} = 1 \implies A = \operatorname{Circ}(0,\dots,1),$$

which is a permutation matrix and is therefore asymmetric. In other words, this matrix moves the last component of a vector to the top while pushing down the other components.

$$\begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & 1\\ 1 & 0 & 0 & \cdots & 0 & 0\\ 0 & 1 & 0 & \cdots & 0 & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & 0 & \cdots & 0 & 0\\ 0 & 0 & 0 & \cdots & 1 & 0\end{pmatrix}\begin{pmatrix} u_1\\ u_2\\ \vdots\\ u_{N-2}\\ u_{N-1}\\ u_N\end{pmatrix} = \begin{pmatrix} u_N\\ u_1\\ u_2\\ \vdots\\ u_{N-2}\\ u_{N-1}\end{pmatrix}.$$

The characteristic polynomial is

$$\begin{aligned}\det(A-\lambda I) &= |\operatorname{Circ}(-\lambda,\dots,1)| = [\text{Expand along first row}]\\ &= (-1)^{N+1}\cdot 1\cdot|\operatorname{tridiag}\begin{pmatrix}0 & 1 & -\lambda\end{pmatrix}| + (-1)^2\cdot(-\lambda)\cdot|\operatorname{tridiag}\begin{pmatrix}1 & -\lambda & 0\end{pmatrix}|\\ &= [\text{Upper \& lower triangular matrices}] = (-1)^{N+1}\cdot 1^{N-1} - \lambda\cdot(-\lambda)^{N-1}\\ &= (-1)^N(\lambda^N - 1).\end{aligned}$$

The roots of this polynomial are $\lambda_k = e^{i2\pi k/N}$, i.e. the N'th roots of unity. Since all eigenvalues are unique and have magnitude one, the method is stable. A faster way to determine stability would be to check the root condition for the characteristic equation of the recurrence relation, which is

$$r = (r^2+1)/2 - a\Delta t\cdot(r^2-1)/(2\Delta x) \implies r = (r^2+1)/2 - (r^2-1)/2 = 1.$$

The only root is $r = 1$, which satisfies the root condition and the method is therefore stable. The reason we want to run at CFL $= 1$ is because then the matrix A is unitary (columns orthogonal to each other) and therefore preserves the norm of the solution when stepping in time. This is a key property of the advection equation and a good method should replicate this behaviour.

A Implementation of problem 4.1

Here is Erik Danielsson's implementation of problem **4.1** (click to go back) which further proves that the presented method is not order one. In figure 2 below, y_1 is computed by expanding the Sturm-Liouville operator

$$\frac{\mathrm{d}}{\mathrm{d}x}(p(x)y'(x)) = p'(x)y'(x) + p(x)y''(x)$$

and then discretizing using standard derivative approximations. For y_2, the method

$$\frac{\mathrm{d}}{\mathrm{d}x}(p(x)y'(x))\bigg|_{x=x_i} \approx \frac{p_{i-1/2}y_{i-1} - (p_{i-1/2} + p_{i+1/2})y_i + p_{i+1/2}y_{i+1}}{h^2}$$

presented in the lecture notes is implemented instead and y_3 is the method given in **4.1**.

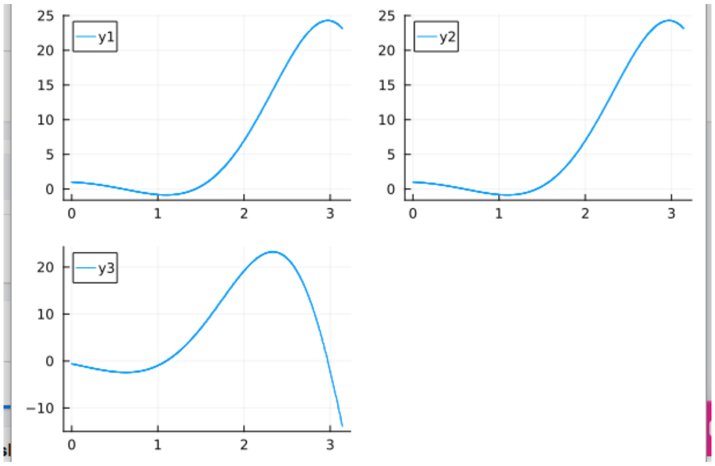

Figure 2: The three different methods for the given problem $p(x) = (1 - 0.8\sin^2 x)$. We see that y_1 and y_2 agree with each other, while y_3 produces a bogus plot.

The implementation was done in Julia and is only 30 lines.

```julia
using Plots

# Numerical approximation of the differential operator
# A = d/dx (p(x) d/dx) where p(x) = 1 - 0.8sin(x)^2

# Expansion and discretisation:
# Ay = p’ y’ + p y’’
# then discretize normally
diff1(p, h, x) = (p(x + h) - p(x - h)) / 2h
diff2(p, h, x) = (p(x + h) - 2p(x) + p(x - h)) / h^2
diffcorr(p, y, h, x) = diff1(p, h, x) * diff1(y, h, x) + p(x) * diff2(y, h, x)

# Direct disretization of A according to lecture notes
diffnorm(p, y, h, x) = (p(x - h / 2) * y(x - h) - (p(x - h / 2) + p(x + h / 2)) * y(x) + p(x + h / 2) * y(x + h)) / h^2

# Method suggested in exercise
diffincorr(p, y, h, x) = (p(x - h) * y(x - h) - 2p(x) * y(x) + p(x + h) * y(x + h)) / h^2

# Example problem
p(x) = 1 - 0.8sin(x)^2
y(x) = exp(x)
N = 1000
h = 1 / (N + 1)
xgrid = range(0, pi, N + 2)

y1 = [diffcorr(p, y, h, x) for x in xgrid]
y2 = [diffnorm(p, y, h, x) for x in xgrid]
y3 = [diffincorr(p, y, h, x) for x in xgrid]

plot(xgrid, [y1, y2, y3], layout=3)
```