

A (a) : $V \in \text{Span}\{W_1, \dots, W_n\}$ $\therefore V = c_1 W_1 + \dots + c_n W_n$

$\because \langle W_i, W_j \rangle = 0$ for $i \neq j$

$$\therefore \langle V, W_j \rangle = c_j \langle W_j, W_j \rangle = c_j \|W_j\|^2, \quad c_j = \frac{\langle V, W_j \rangle}{\|W_j\|^2} \text{ for } j \in \{1, \dots, n\}$$

$$V = c_1 W_1 + \dots + c_n W_n = \sum_{j=1}^n \frac{\langle V, W_j \rangle}{\|W_j\|^2} W_j \quad \square$$

(b) i. N is the dimension, there can be at most N pairwise orthogonal vectors to form a basis. Each iteration spans with a new vector.

$$\text{ii. When } n=1, \quad p_1 = r_1 - \frac{\langle r_1, p_0 \rangle_A}{\|p_0\|_A^2} p_0$$

$$\langle p_1, p_0 \rangle_A = \langle r_1, p_0 \rangle_A - \frac{\langle r_1, p_0 \rangle_A}{\|p_0\|_A^2} \langle p_0, p_0 \rangle_A = \langle r_1, p_0 \rangle_A - \langle r_1, p_0 \rangle_A = 0$$

Suppose $\langle p_n, p_j \rangle_A = 0$ for $0 \leq j < n \leq n^*-1$ holds for all n up to $n=k-1$

$$\text{then when } n=k, \quad p_k = r_k - \sum_{j=0}^{k-1} \frac{\langle r_k, p_j \rangle_A}{\|p_j\|_A^2} p_j$$

$$\begin{aligned} \langle p_k, p_i \rangle_A &= \langle r_k, p_i \rangle_A - \sum_{j=0}^{k-1} \frac{\langle r_k, p_j \rangle_A}{\|p_j\|_A^2} \langle p_j, p_i \rangle_A \\ &= \langle r_k, p_i \rangle_A - \frac{\langle r_k, p_i \rangle_A}{\|p_i\|_A^2} \langle p_i, p_i \rangle_A - \sum_{\substack{j < k \\ j \neq i}} \frac{\langle r_k, p_j \rangle_A}{\|p_j\|_A^2} \langle p_j, p_i \rangle_A \\ &= - \sum_{\substack{0 \leq j < k \\ j \neq i}} \frac{\langle r_k, p_j \rangle_A}{\|p_j\|_A^2} \langle p_j, p_i \rangle_A \quad \text{for } 0 \leq i < k \leq n^*-1 \end{aligned}$$

$\therefore \langle p_j, p_i \rangle_A = 0$ for $0 \leq i, j < k, i \neq j$ $\therefore \langle p_k, p_i \rangle_A = 0$ for $0 \leq i < k \leq n^*-1$ \square

(c) i. $\because \{\phi_1, \dots, \phi_N\}$ is a basis for \mathbb{R}^N , $V, W \in \mathbb{R}^N$

$$\therefore \text{Let } V = a_1 \phi_1 + \dots + a_N \phi_N \quad W = b_1 \phi_1 + \dots + b_N \phi_N$$

$$\therefore A\phi_n = \lambda_n \phi_n \quad \therefore AV = a_1 \lambda_1 \phi_1 + \dots + a_N \lambda_N \phi_N$$

$$\therefore \langle \phi_n, \phi_j \rangle = \delta_{nj} \quad \therefore \langle AV, W \rangle = \sum_{n=1}^N \lambda_n a_n b_n, \quad \langle V, \phi_n \rangle = a_n, \quad \langle \phi_n, W \rangle = b_n$$

$$\langle AV, W \rangle = \sum_{n=1}^N \lambda_n \langle V, \phi_n \rangle \langle \phi_n, W \rangle \quad \square$$

ii. $\because A$ is symmetric positive definite. $\therefore \lambda_n > 0$ for $1 \leq n \leq N$

iii. By i, let $V = a_1 \phi_1 + \dots + a_N \phi_N$, then $\langle AV, V \rangle = \sum_{n=1}^N \lambda_n a_n^2$

$$\|V\|^2 = \langle V, V \rangle = \sum_{n=1}^N a_n^2$$

$$\therefore \lambda_1 \leq \dots \leq \lambda_N \quad \therefore \sum_{n=1}^N \lambda_n a_n^2 \leq \sum_{n=1}^N \lambda_n a_n^2 \leq \sum_{n=1}^N \lambda_N a_n^2, \quad \lambda_N \|V\|^2 \leq \langle AV, V \rangle \leq \lambda_N \|V\|^2 \quad \square$$

$$\text{iv. } AV = a_1 \lambda_1 \phi_1 + \dots + a_N \lambda_N \phi_N, \quad \|AV\|^2 = \langle AV, AV \rangle = \sum_{n=1}^N a_n^2 \lambda_n^2 \leq \sum_{n=1}^N a_n^2 \lambda_N^2 = \lambda_N^2 \|V\|^2$$

$$\therefore \|AV\| \leq \lambda_N \|V\|$$

(d) From LeVeque 2007 page 87:

$$\begin{aligned}
 w_n &= Ap_n & r_{n+1} &= r_n - \alpha_n w_n \\
 p_n &= r_n + \beta_{n-1} p_{n-1} & r_n &= p_n - \beta_{n-1} p_{n-1} \\
 p_{n+1} &= r_{n+1} + \beta_n p_n \\
 &= r_n - \alpha_n w_n + \beta_n p_n \\
 &= p_n - \beta_{n-1} p_{n-1} - \alpha_n A p_n + \beta_n p_n \\
 &= (I + \beta_n) p_n - \alpha_n A p_n - \beta_{n-1} p_{n-1} \quad \text{for } 1 \leq n \leq N-2 \quad \square
 \end{aligned}$$

(e) $A \in \mathbb{R}^{N \times N}$ is non-singular

By Cayley-Hamilton Thm, $p(\lambda) = \det(\lambda I_N - A)$, $p(A) = 0$

$$p(A) = A^N + C_{N-1} A^{N-1} + \dots + C_1 A + (-1)^N \det(A) I_N = 0$$

where C 's are given by the elementary symmetric polynomials of the eigenvalues of A

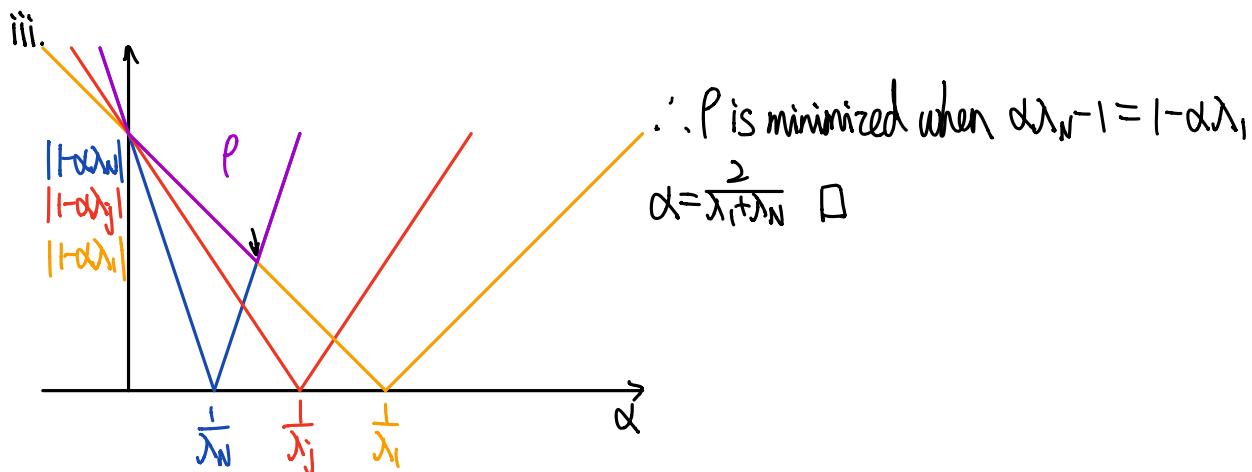
$$\therefore A^N = -C_{N-1} A^{N-1} - \dots - C_1 A - (-1)^N \det(A) I_N \quad \square$$

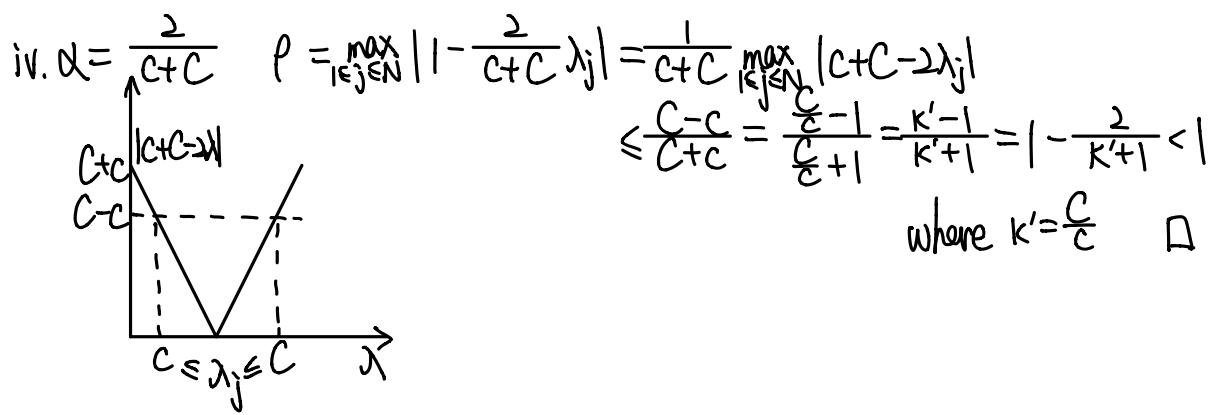
$$\begin{aligned}
 (\text{f}) \text{i. } e_{n+1} &= u_{n+1} - u = u_n - u + \alpha(f - Au_n) - \alpha \underbrace{(f - Au)}_0 \\
 &= u_n - u - \alpha A(u_n + \alpha A u) = (I - \alpha A)(u_n - u) = (I - \alpha A)e_n
 \end{aligned}$$

$$\begin{aligned}
 \text{ii. By (c) iv, } \|e_{n+1}\| &= \|(I - \alpha A)e_n\| \leq \underbrace{\max_{1 \leq j \leq N} |\lambda_j(I - \alpha A)|}_{\text{largest eigenvalue of } (I - \alpha A)} \|e_n\|
 \end{aligned}$$

For eigenvectors $\{v_1, \dots, v_N\}$ of A s.t. $A v_j = \lambda_j v_j$, $(I - \alpha A)v_j = v_j - \alpha \lambda_j v_j = (1 - \alpha \lambda_j)v_j$

\therefore The largest eigenvalue of $(I - \alpha A)$ is $\rho = \max_{1 \leq j \leq N} |1 - \alpha \lambda_j| \quad \square$





(g) $q_n = \frac{r_n}{\|r_n\|}$ for $0 \leq n \leq n^*-1$

$$p_0 = r_0 = f - Ax_0$$

$$w_n = Ap_n \quad r_{n+1} = r_n - \alpha_n w_n \quad p_n = r_n + \beta_{n-1} p_{n-1} \quad p_n = \frac{r_n^\top r_{n+1}}{r_n^\top r_n} = \frac{\|r_{n+1}\|^2}{\|r_n\|^2}$$

i. $r_1 = r_0 - \alpha_0 w_0 = r_0 - \alpha_0 A p_0 = r_0 - \alpha_0 A r_0 \quad \square$

ii. $r_{n+1} = r_n - \alpha_n w_n = r_n - \alpha_n A p_n \quad A p_n = \frac{r_n - r_{n+1}}{\alpha_n}$

$$= r_n - \alpha_n A(r_n + \beta_{n-1} p_{n-1}) = r_n - \alpha_n A r_n - \alpha_n \beta_{n-1} A p_{n-1}$$

$$= r_n - \alpha_n A r_n - \alpha_n \beta_{n-1} \frac{r_{n-1} - r_n}{\alpha_{n-1}} = r_n - \alpha_n A r_n + \frac{\alpha_n \beta_{n-1}}{\alpha_{n-1}} (r_n - r_{n-1}) \text{ for } 1 \leq n \leq n^*-1$$

iii. $x_0 = \frac{1}{\alpha_0} \quad x_n = \frac{1}{\alpha_n} + \frac{\beta_{n-1}}{\alpha_{n-1}} \text{ for } 1 \leq n \leq n^*-1 \quad \delta_n = \frac{\beta_{n-1}}{\alpha_n}$

$$\overline{\beta}_n = \frac{\|r_{n+1}\|}{\|r_n\|} \quad \text{From i, } Ar_0 = \frac{r_0 - r_1}{\alpha_0}$$

$$Aq_0 = \frac{Ar_0}{\|r_0\|} = \frac{1}{\alpha_0} \frac{r_0}{\|r_0\|} - \frac{1}{\alpha_0} \frac{r_1}{\|r_0\|} = \frac{1}{\alpha_0} \frac{r_0}{\|r_0\|} - \frac{\overline{\beta}_0}{\alpha_0} \frac{r_1}{\|r_0\|} = x_0 q_0 - \delta_0 q_1$$

From ii, $Ar_n = \frac{1}{\alpha_n} (r_n - r_{n+1}) + \frac{\beta_{n-1}}{\alpha_{n-1}} (r_n - r_{n-1})$

$$Aq_n = \frac{Ar_n}{\|r_n\|} = \frac{1}{\alpha_n} \frac{r_n}{\|r_n\|} - \frac{1}{\alpha_n} \frac{r_{n+1}}{\|r_n\|} + \frac{\beta_{n-1}}{\alpha_{n-1}} \frac{r_n}{\|r_n\|} - \frac{\beta_{n-1}}{\alpha_{n-1}} \frac{r_{n-1}}{\|r_n\|}$$

$$= \left(\frac{1}{\alpha_n} + \frac{\beta_{n-1}}{\alpha_{n-1}} \right) \frac{r_n}{\|r_n\|} - \frac{\overline{\beta}_n}{\alpha_n} \frac{r_{n+1}}{\|r_n\|} - \frac{\overline{\beta}_{n-1}}{\alpha_{n-1}} \frac{r_{n-1}}{\|r_n\|} = x_n q_n - \delta_n q_{n+1} - \delta_{n-1} q_{n-1} \text{ for } 1 \leq n \leq n^*-1 \quad \square$$

iv. $Q_n = [q_0 \cdots q_{n-1}] \in \mathbb{R}^{N \times n}$

$$T_n = \begin{bmatrix} x_0 & -\delta_0 \\ -\delta_0 & x_1 & -\delta_1 \\ & \ddots & \ddots & \ddots \\ & & -\delta_{n-3} & x_{n-2} & -\delta_{n-2} \\ & & & -\delta_{n-2} & x_{n-1} \end{bmatrix} \in \mathbb{R}^{n \times n} \quad e_n = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \in \mathbb{R}^n$$

$$Q_n T_n = [(x_0 q_0 - \delta_0 q_1) \quad (-\delta_0 q_0 + x_1 q_1 - \delta_1 q_2) \quad \cdots \quad (-\delta_{n-3} q_{n-3} + x_{n-2} q_{n-2} - \delta_{n-2} q_{n-1}) \quad (-\delta_{n-2} q_{n-2} + x_{n-1} q_{n-1})]$$

$$-\delta_{n-1} q_n e_n^\top = [0 \cdots 0 \quad -\delta_{n-1} q_n]$$

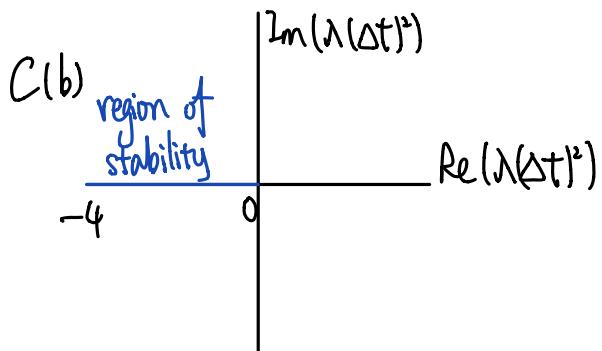
$$\therefore \text{By iii, } A Q_n = [(\gamma_0 q_0 - \delta_0 q_1) \quad (-\delta_0 q_0 + \gamma_1 q_1 - \delta_1 q_2) \quad \cdots \quad (-\delta_{n-3} q_{n-3} + \gamma_{n-2} q_{n-2} - \delta_{n-2} q_{n-1}) \quad (-\delta_{n-2} q_{n-2} + \gamma_{n-1} q_{n-1} - \delta_{n-1} q_n)] \\ = Q_n T_n - \delta_{n-1} q_n e_n^T \quad \square$$

v. $\because \{q_0, \dots, q_{n-1}\}$ is an orthonormal basis $\therefore q_i^T \cdot q_j = \langle q_i, q_j \rangle = \delta_{ij} \leftarrow \text{Dirac}$

$$Q_n^T Q_n = \begin{bmatrix} q_0^T q_0 & \cdots & q_0^T q_{n-1} \\ \vdots & \ddots & \vdots \\ q_{n-1}^T q_0 & \cdots & q_{n-1}^T q_{n-1} \end{bmatrix} = \begin{bmatrix} \delta_{00} & \cdots & \delta_{0,n-1} \\ \vdots & \ddots & \vdots \\ \delta_{n-1,0} & \cdots & \delta_{n-1,n-1} \end{bmatrix} = I_n$$

$$Q_n^T \delta_{n-1} q_n e_n^T = \begin{bmatrix} 0 & \cdots & 0 & \vdots \\ \vdots & & \ddots & \delta_{n-1} q_n^T q_n \end{bmatrix} = 0$$

$$Q_n^T A Q_n = Q_n^T Q_n T_n - Q_n^T \delta_{n-1} q_n e_n^T = T_n \quad \square$$



Math 714 HW 2

Diya Yang

A See handwritten notes.

B (10 pts) Consider the function

$$f(x) = e^{-400(x-0.5)^2}$$

for $x \in [0, 1]$. Sample it on a grid $x_j = jh$ with $h = 1/N$ and $0 \leq j \leq N$, for some N to be determined. Consider the linear interpolant of f computed from the $N + 1$ samples $f(x_j)$. Numerically, find the smallest value of N such that f differs from its linear interpolant by at most 10^{-2} in the uniform norm. [Hint: Matlab has interp1.m for 1 D linear interpolation.]

Solution:

Here's the link to codes on GitHub.

I linearly interpolated the function $f(x)$ with grid points evenly spaced between 0 and 1 with distance of $\frac{1}{N}$, and found the largest error of the interpolant among query points that are evenly spaced between 0 and 1 with distance of $\frac{1}{1000N}$. The smallest value of N such that f differs from its linear interpolant by at most 10^{-2} in the uniform norm (based on these query points) is $N = 100$.

C Consider the 2D wave equation

$$u_{tt} = \Delta u, \quad 0 \leq x, y \leq 1$$

with homogeneous Dirichlet boundary conditions. Fix the initial conditions to be

$$u(x, y, 0) = 0, \quad u_t(x, y, 0) = f(x)f(y)$$

where f was defined in problem B. Consider a spatial grid $\mathbf{x}_j = (x_{j_1}, y_{j_2}) = (j_1\Delta x, j_2\Delta x)$ with Δx small enough to resolve the initial condition, in the sense of problem B. (I.e., take Δx less than $h = 1/N$, where the critical N was found in problem B.)

(a) (30pts) Implement and test the "simplest" numerical method, which uses the 3 -point formula for the second derivative in time, and the 5 -point Laplacian at time t_n . It results in a two-step method. Explain how you initialize your scheme. Show a log-log plot of the error vs. the grid spacing Δx , and check from this plot that your method is second-order accurate.

Solution:

Let $x_i = i\Delta x$, $y_j = j\Delta x$, $t_n = n\Delta t$ where $i, j = 0, 1, \dots, M$, $n = 0, 1, \dots, T$, $\Delta x = \frac{1}{M}$.

Using the 3 -point formula for the second derivative in time, and the 5 -point Laplacian at time t_n , we have

$$\frac{1}{(\Delta t)^2}(U_{i,j}^{n-1} - 2U_{i,j}^n + U_{i,j}^{n+1}) = \frac{1}{(\Delta x)^2}(U_{i-1,j}^n + U_{i+1,j}^n + U_{i,j-1}^n + U_{i,j+1}^n - 4U_{i,j}^n) \quad (1)$$

where $U_{i,j}^n$ is an approximation for $u(x_i, y_j, t_n)$.

Let $r = \left(\frac{\Delta t}{\Delta x}\right)^2$ and $U^n = \begin{pmatrix} U_{1,1}^n \\ \vdots \\ U_{1,M-1}^n \\ U_{2,1}^n \\ \vdots \\ U_{2,M-1}^n \\ \vdots \\ U_{M-1,1}^n \\ \vdots \\ U_{M-1,M-1}^n \end{pmatrix}_{(M-1)^2}$, then

M^{th} column
 \downarrow

$U^{n-1} + U^{n+1} = \begin{bmatrix} 2 - 4r & r & \cdots & r & & & & \\ r & 2 - 4r & r & \cdots & r & & & \\ \vdots & \ddots & \ddots & \ddots & & \ddots & & \\ r & & \ddots & \ddots & \ddots & & & \\ & \ddots & & \ddots & \ddots & & & \\ & & & \ddots & \ddots & \ddots & & \\ r & \cdots & r & 2 - 4r & r & & & \\ r & \cdots & r & 2 - 4r & r & & & \end{bmatrix}_{(M-1)^2 \times (M-1)^2}$

$U^n + \text{boundary conditions}$ (2)

since we have homogeneous Dirichlet boundary conditions. Call this matrix A .

Since $u_t(x, y, 0) = f(x)f(y)$, we can use ghost nodes such that

$$\frac{U_{i,j}^1 - U_{i,j}^{-1}}{2\Delta t} = f(x_i)f(y_j) \quad (3)$$

We are also using U^T as the ghost node at U^{T+1} as t only has boundary conditions at one side. Therefore,

$$\begin{bmatrix} -A & 2I & & \\ I & -A & I & \\ & \ddots & \ddots & \ddots & \\ & & I & -A & I \\ & & & I & -A + I \end{bmatrix} \begin{bmatrix} U^0 \\ U^1 \\ \vdots \\ U^T \end{bmatrix} = \begin{bmatrix} F \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \text{where } F = 2\Delta t \begin{pmatrix} f(x_1)f(y_1) \\ \vdots \\ f(x_1)f(y_{M-1}) \\ \vdots \\ f(x_{M-1})f(y_1) \\ \vdots \\ f(x_{M-1})f(y_{M-1}) \end{pmatrix} \quad (4)$$

Implementing this scheme and compare with the results from a finer grid, we have

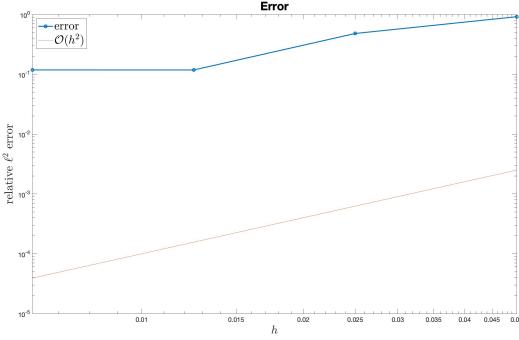


Figure 1: log-log plot of the ℓ^2 norm of the error vs. h

Here's the link to codes on GitHub.

- (b) (5 pts) Consider the ODE $y''(t) = \lambda y$, and the 3 -point rule for y'' as a two-step explicit time integrator. Find the region of stability of this ODE solver in terms of $\lambda(\Delta t)^2$, and plot it in the complex plane.

Solution:

$$\frac{y_{n-1} - 2y_n + y_{n+1}}{(\Delta t)^2} = \lambda y_n \quad (5)$$

$$y_{n+1} - (2 + \lambda(\Delta t)^2)y_n + y_{n-1} = 0 \quad (6)$$

$$\rho^2 - (2 + \lambda(\Delta t)^2)\rho + 1 = 0 \quad (7)$$

$$\rho = \frac{2 + \lambda(\Delta t)^2 \pm \sqrt{(2 + \lambda(\Delta t)^2)^2 - 4}}{2} \quad (8)$$

Let $x = 2 + \lambda(\Delta t)^2 \in \mathbb{C}$, then

$$|\rho|^2 = \frac{|x|^2 + |x^2 - 4| \pm 2\operatorname{Re}(\bar{x}\sqrt{x^2 - 4})}{4} \quad (9)$$

Since $|x|^2 + |x^2 - 4| \geq 4$ and $|x|^2 + |x^2 - 4| = 4$ when $x \in \mathbb{R}$, the system is stable when $x \in \mathbb{R}$ and $\operatorname{Re}(\bar{x}\sqrt{x^2 - 4}) = 0$.

Therefore, $x \in [-2, 2] \subset \mathbb{R}$, $\lambda(\Delta t)^2 = x - 2 \in [-4, 0] \subset \mathbb{R}$.

Plot: see handwritten notes.

- (c) (5pts) From your answer to (b), and your knowledge of the spectrum of the discrete Laplacian, perform the “method of lines” stability analysis for the method in (a). What CFL condition does this analysis result in?
- (d) (5pts) Perform the von Neumann stability analysis for the method in (a), and check if the resulting CFL condition agrees with what you found in the previous question. [Hint: since this is a 2D problem, a plane wave is $\exp(ik_1 j_1 \Delta x) \exp(ik_2 j_2 \Delta x)$.]
- (e) (5pts) Find the modified equation that corresponds to the numerical method in (a). Solve it via Fourier series, and comment on the physics of the extra terms. Are they dissipative, dispersive, or something else?
- D (Bonus 20 pts) Formulate and prove one extension of the Lax equivalence theorem to the case of linear ODE with two time derivatives.

Solution:

Consider the second-order linear ODE:

$$u''(t) + au'(t) + bu(t) + c(t) = 0 \quad (10)$$

with boundary conditions $u(0) = u_0$ and $u'(0) = u'_0$, then

$$\frac{U_{n-1} - 2U_n + U_{n+1}}{(\Delta t)^2} + \frac{a}{2\Delta t}(U_{n+1} - U_{n-1}) + bU_n + c_n = 0 \quad (11)$$

$$\frac{2 + a\Delta t}{2(\Delta t)^2}U_{n+1} = (\frac{2}{(\Delta t)^2} - b)U_n + \frac{a\Delta t - 2}{2(\Delta t)^2}U_{n-1} - c_n \quad (12)$$

$$U_{n+1} = A(\Delta t)U_n + B(\Delta t)U_{n-1} + C_n(\Delta t) \quad (13)$$

$$u_{n+1} = Au_n + Bu_{n-1} + C_n + \Delta t\tau_n \quad (14)$$

$$E_n = U_n - u_n \quad (15)$$

$$E_{n+1} = AE_n + BE_{n-1} - \Delta t\tau_n \quad (16)$$

$$E_N = \sum_{k=0}^{\lfloor (N-1)/2 \rfloor} \binom{N-k-1}{k} A^{N-2k-1} B^k E_1 + \sum_{k=0}^{\lfloor (N-2)/2 \rfloor} \binom{N-k-2}{k} A^{N-2k-2} B^{k+1} E_0 \quad (17)$$

$$+ \Delta t \sum_{n=0}^{N-1} \sum_{k=0}^{\lfloor (N-n-1)/2 \rfloor} \binom{N-n-k-1}{k} A^{N-n-2k-1} B^k \tau_n \quad (18)$$

Since we have fixed boundary, $E_0 = 0$ and $E_1 \rightarrow 0$ as $\Delta t \rightarrow 0$. $\tau_n \rightarrow 0$ as $\Delta t \rightarrow 0$ because of consistency.

Weak Stability \Rightarrow Convergence:

Weak stability states that for any $T > 0$, there exists $C_T > 0$ such that $\left\| \sum_{k=0}^{\lfloor (N-1)/2 \rfloor} \binom{N-k-1}{k} A^{N-2k-1} B^k \right\| \leq C_T$ for any Δt such that $n\Delta t \leq T$. Similarly, $\left\| \sum_{k=0}^{\lfloor (N-n-1)/2 \rfloor} \binom{N-n-k-1}{k} A^{N-n-2k-1} B^k \right\|$ are also bounded by C'_T . Therefore,

$$E_N \leq C_T E_1 + \Delta t N C'_T \max_n \tau_n \rightarrow 0 \text{ as } \Delta t \rightarrow 0 \quad (19)$$

This system is convergent.

Not Weakly Stable \Rightarrow Not Convergent:

If this system is not weakly stable, then there exists $T > 0$ such that for any $C_T > 0$ there exists Δt satisfying $n\Delta t \leq T$ such that $\left\| \sum_{k=0}^{\lfloor (N-1)/2 \rfloor} \binom{N-k-1}{k} A^{N-2k-1} B^k \right\| > C_T$. Take $C_T > \frac{1}{\|E_1\|}$, then $\sum_{k=0}^{\lfloor (N-1)/2 \rfloor} \binom{N-k-1}{k} A^{N-2k-1} B^k E_1$ does not converge to 0. Similar for the terms with τ_n , and these terms cannot be cancelled. Therefore, this system is not convergent if it is not weakly stable. \square