

Numerical Mathematics II for Engineers - Homework 4

Group 11

Hetvi Chaniyara, Nelleke Kortleven, Hande Pamuksuz

November 23, 2025

Exercise 5.1

(a)

The infinity norm of a vector v , $\|v\|_\infty$ is defined as the largest absolute value among all the components v_i of the vector v . We show that the norm induced by $\|\cdot\|_\infty$ satisfies:

1. $\|A\|_\infty \leq \max_{i \in \{1, \dots, N\}} \sum_{j=1}^N |a_{ij}| \quad \text{for all } A \in \mathbb{R}^{N \times N};$
2. $\|A\|_\infty \geq \max_{i \in \{1, \dots, N\}} \sum_{j=1}^N |a_{ij}| \quad \text{for all } A \in \mathbb{R}^{N \times N}.$

(1)

$$\begin{aligned} \|A\|_\infty &= \sup_{v \neq 0} \frac{\|Av\|_\infty}{\|v\|_\infty} \\ &= \sup_{\|v\|_\infty=1} \|Av\|_\infty && [\text{Definition of the induced norm, restricted to unit vectors}] \\ &= \sup_{\|v\|_\infty=1} \left(\max_i \left| \sum_{j=1}^N a_{ij} v_j \right| \right) && [\text{Definition of the vector } \infty\text{-norm: } \|x\|_\infty = \max_i |x_i|] \\ &\leq \sup_{\|v\|_\infty=1} \left(\max_i \sum_{j=1}^N |a_{ij}| |v_j| \right) && [\text{Application of the Triangle Inequality: } |\sum c_j| \leq \sum |c_j|] \\ &\leq \sup_{\|v\|_\infty=1} \left(\max_i \sum_{j=1}^N |a_{ij}| \cdot 1 \right) && [\text{Since } \|v\|_\infty = 1, |v_j| \leq 1 \text{ for all } j] \\ &= \max_i \sum_{j=1}^N |a_{ij}| && [\sup_{\|v\|_\infty=1} (\text{Constant}) = \text{Constant}]. \end{aligned}$$

(2)

Let i^* be the row where the maximum is attained:

$$\max_i \sum_{j=1}^N |a_{ij}| = \sum_{j=1}^N |a_{i^*j}|.$$

We construct a specific test vector v^* where:

$$v_j^* = \begin{cases} 1 & \text{if } a_{i^*j} \geq 0 \\ -1 & \text{if } a_{i^*j} < 0 \end{cases} = \operatorname{sgn}(a_{i^*j}).$$

Then $\|v^*\|_\infty = 1$ and the i^* -th component of Av^* is:

$$(Av^*)_{i^*} = \sum_{j=1}^N a_{i^*j} v_j^* = \sum_{j=1}^N a_{i^*j} \cdot \operatorname{sgn}(a_{i^*j}) = \sum_{j=1}^N |a_{i^*j}|$$

The ∞ -norm of the resulting vector is at least this component:

$$\|Av^*\|_\infty \geq |(Av^*)_{i^*}| = \sum_{j=1}^N |a_{i^*j}| = \max_i \sum_{j=1}^N |a_{ij}|.$$

Since $\|v^*\|_\infty = 1$, the definition of the induced norm is:

$$\|A\|_\infty = \sup_{v \neq 0} \frac{\|Av\|_\infty}{\|v\|_\infty} \geq \frac{\|Av^*\|_\infty}{\|v^*\|_\infty} = \|Av^*\|_\infty \geq \max_i \sum_{j=1}^N |a_{ij}|.$$

Conclusion:

$$\|A\|_\infty = \max_{i \in \{1, \dots, N\}} \sum_{j=1}^N |a_{ij}| \quad \text{for all } A \in \mathbb{R}^{N \times N}.$$

(b)

Definition from Lecture 08: an FD scheme is **stable** w.r.t. $(\|\cdot\|_h)_{h>0}$ if $L_h \in \mathbb{R}^{(N-1) \times (N-1)}$ is invertible and if there exists $C > 0$ with $\|L_h^{-1}\|_h \leq C < \infty$ for eventually all $h > 0$.

Given:

- The inequality $\|A\|_2 \leq \sqrt{N} \|A\|_\infty$ holds for all $A \in \mathbb{R}^{N \times N}$;
- The scheme is stable w.r.t. $\|\cdot\|_{\infty,h}$, meaning: $\|L_h^{-1}\|_{\infty,h} \leq C < \infty$ for all $h > 0$, where C is independent of h ;
- We have $h = \frac{1}{N+1}$, so $N = \frac{1}{h} - 1$.

Applying the given inequality to L_h^{-1} gives:

$$|||L_h^{-1}|||_{2,h} \leq \sqrt{N} \cdot |||L_h^{-1}|||_{\infty,h} \leq \sqrt{N} \cdot C.$$

Since $N = \frac{1}{h} - 1$, we have:

$$|||L_h^{-1}|||_{2,h} \leq C\sqrt{N} = C\sqrt{\frac{1}{h} - 1} \leq C\sqrt{\frac{1}{h}} = \frac{C}{\sqrt{h}}.$$

As $h \rightarrow 0$, this grows unboundedly.

For stability w.r.t. $|||\cdot|||_{2,h}$, we need $|||L_h^{-1}|||_{2,h} \leq C' < \infty$ where C' is independent of h . However, the best bound we can obtain from the infinity norm stability is $C\sqrt{N} \sim Ch^{-1/2}$, which depends on h and is unbounded as $h \rightarrow 0$.

Therefore, stability in the infinity norm does not imply stability in the 2-norm for finite difference schemes.

Exercise 5.2

(a)

The assumption is

$$|u - u_h|_\infty \leq c h^p$$

Applying this to mesh size at $h/2$, we get the following:

$$|u - u_{h/2}|_\infty \leq c h^p / 2^p$$

The maximum difference between u_h and $u_{h/2}$ can also be written as:

$$|(u - u_{h/2}) - (u - u_h)|_\infty$$

Based on the assumption, the righthand side can also be completed:

$$|u_h - u_{h/2}| \leq c h^p + \frac{c h^p}{2^p} = c \left(1 + \frac{1}{2^p}\right) h^p$$

We can define $\tilde{c} = c \left(1 + \frac{1}{2^p}\right)$. Then we reach equation (1);

$$\|u_h - u_{h/2}\|_\infty \leq \tilde{c} h^p \quad \square$$

(b)

The assumption is

$$|u - u_h|_\infty \leq c h^p$$

We can copy the logic from part (a) and write this for mesh sizes h , $h/2$ and $h/4$ for experimentation.

Looking at the maximum difference between h and $h/2$, and $h/2$ and $h/4$, we can calculate a ratio of change for each time we halve h . First:

$$u_h - u_{h/2} = (u - u_{h/2}) - (u - u_h) \approx c(h/2)^p - c h^p = c h^p ((1/2)^p - 1)$$

and:

$$u_{h/2} - u_{h/4} = (u - u_{h/4}) - (u - u_{h/2}) \approx c(h/4)^p - c(h/2)^p = c h^p ((1/4)^p - (1/2)^p)$$

The ratio R is:

$$R = \frac{D_1}{D_2} \approx \frac{|c|h^p \left(1 - \frac{1}{2^p}\right)}{|c|\frac{h^p}{2^p} \left(1 - \frac{1}{2^p}\right)} = 2^p.$$

The ratio of the differences between successive solutions is 2^p , where p is the unknown convergence order.

Here we got the experimental order of convergence as

$$p = \log_2(R)$$

Exercise 5.3

$$-u''(x) + g(x)u(x) = f(x), \quad x \in (a, b) \quad (1)$$

$$\alpha u(a) + \beta u'(a) = 1, \quad \delta u(b) + \gamma u'(b) = 1 \quad (2)$$

(a)

From the questions we have the α and β as coefficients in the left boundary condition and δ and γ as coefficients in the right boundary condition. We categorize the four test cases into Dirichlet, Neumann and Robin.

- A: $\alpha = 0$, $\beta = 1$, $\delta = 0$, $\gamma = 1$; The left boundary is defined by the derivative $u'(a)$ and the right boundary is also defined by the derivative $u'(b)$ thereby making this the **pure Neumann Boundary condition**.
- B: $\alpha = 1$, $\beta = 0$, $\delta = 0.5$, $\gamma = 0$; The left and right boundaries are defined by some constant $\alpha u(a)$ and $\delta u(b)$ as the derivative have been set to zero. This is the **pure Dirichlet Boundary condition**.

- C: $\alpha = 0, \beta = 1, \delta = 1, \gamma = 2$; The left boundary condition only has the derivative term making this a Neumann boundary condition. The right boundary condition has both coefficients as non-zero, making this the Robin boundary condition. This makes the overall boundary condition a **mixed Neumann-Robin Boundary condition**.
- D: $\alpha = 1, \beta = -\frac{1}{3}, \delta = -\frac{11}{9}, \gamma = 1$; None of the coefficients are zero, making both left and right boundary conditions as Robin. The overall boundary condition would be a **pure Robin Boundary condition**.

The solvability of these test-problems is defined by the Fredholm Alternative Theorem. The problem is uniquely solvable unless the associated homogeneous problem ($f(x) = 0$ and boundary values = 0) has non-trivial solutions.

Case A:

Since we have $g(x) = x * 0$, the first equation simplifies to $-u''(x) = f(x)$ with the pure Neumann boundary conditions $u'(a) = 1$ and $u'(b) = 1$. The homogeneous version, $-u''(x) = 0$ with $u'(x)(a) = 0$ and $u'(x)(b) = 0$, has the non-trivial solution $u(x) = C$, where C is any constant. A solution exists only if the inputs satisfy the Compatibility Condition (aka. flux balance):

$$\int_0^1 f(x) dx = u'(1) - u'(0)$$

The total source/sink (LHS) and net flux (RHS) are calculated using the prescribed values $f(x) = x^2 + 2x - 1$, $u'(0) = 1$, and $u'(1) = 1$:

$$\begin{aligned} \text{Total Source} &= \int_0^1 (x^2 + 2x - 1) dx = \left[\frac{x^3}{3} + x^2 - x \right]_0^1 = \frac{1}{3} \\ \text{Net Flux} &= u'(1) - u'(0) = 1 - 1 = 0 \end{aligned}$$

Since $\frac{1}{3} \neq 0$, the flux balance is violated. Therefore, the linear problem in case A is not solvable (no solution exists).

Cases B,C,D:

The presence of at least one dirichlet or robin boundary condition, combined with $g(x) \geq 0$ in the interval $[0,1]$, ensures the associated homogeneous problem has only the trivial solution ($u \equiv 0$). This guarantees a unique solution.

(b)

The system is the $(N+2) \times (N+2)$ linear system and $h = \frac{b-a}{N+1}$ is the step size.

Discretize Interior Nodes (Row $i = 1, \dots, N$)

We can discretize the interior nodes by using the centeral difference quotient for $u''(x_i)$:

$$u''(x_i) \approx \frac{u_{i+1} - 2u_i + u_{i-1}}{h^2}$$

Substituting this approximation into the PDE, $-u''(x_i) + g(x_i)u(x_i) = f(x_i)$:

$$-\frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} + g(x_i)u_i = f(x_i)$$

Multiplying the entire equation by h^2 yields the final row equation:

$$-u_{i-1} + (2 + h^2 g(x_i))u_i - u_{i+1} = h^2 f(x_i)$$

Now we need to discretize the boundary conditions using a forward/backward difference coefficient.

Left Boundary (Row $i = 0$)

The Left BC is $\alpha u(a) + \beta u'(a) = 1$.

Using the forward difference quotient for $u'(a)$:

$$u'(a) \approx \frac{u_1 - u_0}{h}$$

Substituting the approximation into the BC and rearranging terms for u_0 and u_1 :

$$\begin{aligned} \alpha u_0 + \beta \left(\frac{u_1 - u_0}{h} \right) &= 1 \\ \left(\alpha - \frac{\beta}{h} \right) u_0 + \left(\frac{\beta}{h} \right) u_1 &= 1 \end{aligned}$$

Right Boundary (Row $i = N + 1$)

The Right BC is $\delta u(b) + \gamma u'(b) = 1$.

Using the backward difference quotient for $u'(b)$:

$$u'(b) \approx \frac{u_{N+1} - u_N}{h}$$

Substituting the approximation into the BC and rearranging terms for u_N and u_{N+1} :

$$\begin{aligned} \delta u_{N+1} + \gamma \left(\frac{u_{N+1} - u_N}{h} \right) &= 1 \\ \left(-\frac{\gamma}{h} \right) u_N + \left(\delta + \frac{\gamma}{h} \right) u_{N+1} &= 1 \end{aligned}$$

The Full Discrete Linear System $\mathbf{A}\mathbf{u} = \mathbf{b}$

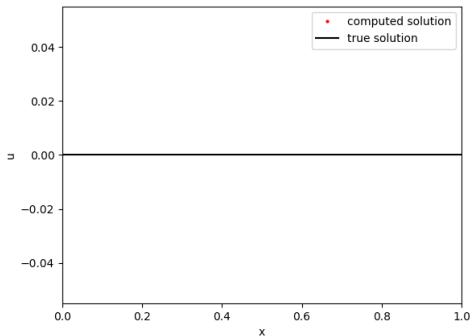
The matrix \mathbf{A} contains the coefficients:

$$\mathbf{A} = \begin{pmatrix} \left(\alpha - \frac{\beta}{h} \right) & \frac{\beta}{h} & 0 & \cdots & 0 & 0 \\ -1 & \left(2 + h^2 g_1 \right) & -1 & \cdots & 0 & 0 \\ 0 & -1 & \left(2 + h^2 g_2 \right) & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \left(2 + h^2 g_N \right) & -1 \\ 0 & 0 & 0 & \cdots & \left(-\frac{\gamma}{h} \right) & \left(\delta + \frac{\gamma}{h} \right) \end{pmatrix}$$

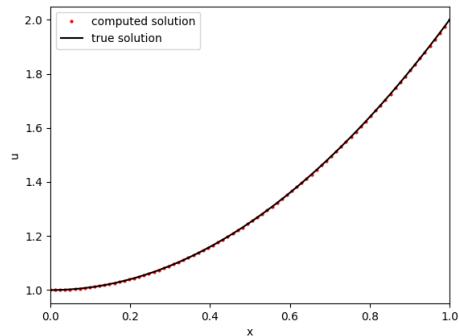
where $g_i = g(x_i)$. The RHS vector \mathbf{b} incorporates $f(x)$ and the fixed boundary values:

$$\mathbf{b} = \begin{pmatrix} 1 \\ h^2 f(x_1) \\ h^2 f(x_2) \\ \vdots \\ h^2 f(x_N) \\ 1 \end{pmatrix}$$

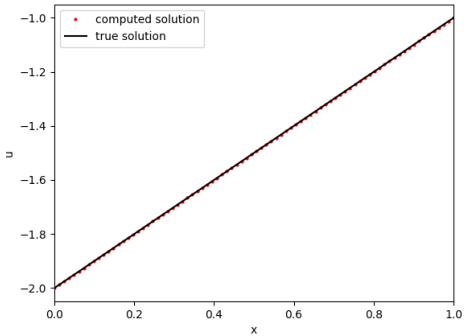
(d)



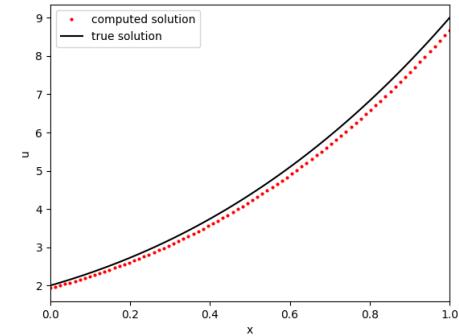
(a) Solution for Case A



(b) Solution for Case B



(c) Solution for Case C



(d) Solution for Case D

Figure 1: Solutions for all test problems

The results across all four test cases are consistent with the predicted behaviour from part (a). Case A produces NaN values as predicted because the problem is a pure Neumann boundary value problem and is not solvable. Cases B and C follow the true solution almost exactly because the exact answers were

simple, low-order curves (parabola and a line). Case D, however, deviates from the true solution curve. This suggests that the overall accuracy is limited by the simpler, less accurate $O(h)$ approximations used at the complex Robin boundaries when the true solution is a cubic curve. The reason Case C still showed a perfect match despite having Robin boundaries is that its exact solution was a straight line, for which the $O(h)$ boundary approximation is exact, eliminating that error source.