MA 214: Introduction to Numerical Analysis

Indian Institute of Technology Bombay

Midsemester Examination

Time: **120 minutes** Marks: **35**

Instructor: **S. Sivaji Ganesh and S. Baskar** Date: **23-02-2023**

Part - 2

Each question carries 4 marks.

12. Evaluate an approximate value of the function $f(x) = \sin x$ at $x = \pi/3$ using $T_4(x)$ about the point $a = 0$. Obtain the remainder $R(\pi/3)$ in terms of some unknown real number ξ. Compute approximately a possible value of ξ by considering the exact value of the function as $f(\pi/3) = 0.866025$.

Answer.

Using the Taylor's theorem, we have

$$\begin{aligned} f(x) &= f(0) + f'(0)x + f''(0)\frac{x^2}{2!} + f'''(0)\frac{x^3}{3!} + f^{(4)}(0)\frac{x^4}{4!} + f^{(5)}(\xi)\frac{x^5}{5!} \\ &= x - \frac{x^3}{3!} + \cos(\xi)\frac{x^5}{5!}. \end{aligned}$$

(A(General formula) + B(expression): 0.5 + 0.5 = 1 mark)

If a student took the reminder as $\sin(\xi)\frac{x^4}{4!}$, but the idea in the following part is correct, then partial marks are given as per the correctness of the following part. ONLY marks for Steps A and B are not given in this case.

where ξ lies between 0 and x. **(C: 0.5 marks)**

Therefore,

$$T_4(x) = x - \frac{x^3}{3!} \quad \text{and} \quad R(x) = \cos(\xi)\frac{x^5}{5!}.$$

Put $x = \pi/3$ in $T_4(x)$ to get an approximate value of $f(\pi/3)$ and is given by

$$f(\pi/3) \approx T_4(\pi/3) \approx 0.8558007815651173.$$

(D: 0.5 marks)

From the remainder term, we get

$$R(\pi/3) \approx 0.010494502221717465 \times \cos(\xi)$$

$$\Big(\text{or} \quad R(\pi/3) \approx \ 0.0501075571162564 \times \sin(\xi)\Big).$$

(E: 0.5 marks)

To obtain values of the unknown ξ, we have to know the exact value of $f(\pi/3)$. Let us take it as $f(\pi/3) = 0.866025$ (rounded to 6 decimal places). Then,

---

$$f(\pi/3) - T_4(\pi/3) = R(\pi/3)$$

$$\implies \quad 0.866025 - 0.855801 = 0.010494502221717465 \times \cos(\xi).$$

$$\Big(\text{ or } \quad \implies 0.866025 - 0.855801 = 0.0501075571162564 \times \sin(\xi).\Big)$$

One of the two equations is enough.

(F : 0.5 marks)

$$\Rightarrow \quad \cos(\xi) \approx 0.9742452018090588.$$

$$\Big(\text{or} \Rightarrow \sin(\xi) \approx 0.204045437919896.\Big)$$

(G: 0.5 marks)

Taking $\cos^{-1}$ (or $\sin^{-1}$) on both sides, we get

$$\xi \approx 0.2274472044318817.$$

$$\Big(\text{or } \xi \approx 0.205488531698947.\Big)$$

(H: 0.5 marks)

---

13. Let $A \in M_n(\mathbb{R})$ be a symmetric and positive definite matrix, and let $L = (l_{ij})$, $i, j = 1, 2, \ldots, n$, be a lower triangular matrix obtained using the Cholesky's factorization method such that $A = LL^T$. Recall that the non-diagonal elements of L are computed using the formula

$$l_{ij} = \left(a_{ij} - \sum_{k=1}^{j-1} l_{ik} l_{jk} \right) \Big/ \, l_{jj},$$

for each $i = 1, 2, \ldots, n$ and for $j = 1, 2, \ldots, i-1$. If N denotes the total number of arithmetic operations involved in computing all the non-diagonal elements in L (assume that the diagonal elements are computed already and hence the operation count of the diagonal elements is NOT included here), then find the constants a, b, c, and d such that

$$N = an^3 + bn^2 + cn + d.$$

Answer.

From the given expression for l_{ij}, we see that there are one division, one subtraction, $j-2$ additions and $j-1$ multiplications involved in the computation.

(A: 0.5 marks)

We therefore consider $j-1$ additions/subtractions and j multiplications/divisions involved in the expression. There are $i-1$ such expression to be computed for the i^{th} row. Therefore, the total number of arithmetic operations involved in the computation of all the non-diagonal elements is

$$\text{Total flops count for non-diagonal elements} = \sum_{i=1}^{n} \sum_{j=1}^{i-1} (j-1) + \sum_{i=1}^{n} \sum_{j=1}^{i-1} j$$

We have combined additions and subtractions, similarly we have also combined multiplications and divisions. A student may write them individually and then write the expressions accordingly. This is acceptable.

(B (Add/Sub) + C(Mult/Div): 0.5 + 0.5 = 1 mark)

We use the following formulae

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}, \quad \sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}.$$

By taking $k = i - 1$, we see that

$$\sum_{i=1}^{n} (i-1) = \sum_{k=1}^{n-1} k = \frac{n(n-1)}{2}.$$

Addition and Subtraction:

$$\sum_{i=1}^{n} \sum_{j=1}^{i-1} (j-1) \;=\; \sum_{i=1}^{n} \left(\sum_{k=1}^{i-2} k \right) = \frac{1}{2} \sum_{i=1}^{n} (i^2 - 3i + 2)$$

(D: 0.5 marks)

Using the above formulae, we get

$$\begin{aligned} \sum_{i=1}^{n} \sum_{j=1}^{i-1} (j-1) &= \frac{1}{2} \left\{ \frac{n(n+1)(2n+1)}{6} - 3\frac{n(n+1)}{2} + 2n \right\} \\ &= \frac{n^3}{6} - \frac{n^2}{2} + \frac{n}{3} \end{aligned}$$

(E: 0.5 marks)

Multiplication and Division:

$$\begin{aligned} \sum_{i=1}^{n} \sum_{j=1}^{i-1} j &= \sum_{i=1}^{n} \frac{i(i-1)}{2} \\ &= \frac{1}{2} \left(\sum_{i=1}^{n} i^2 - \sum_{i=1}^{n} i \right). \end{aligned}$$

(F: 0.5 marks)

Using the above formulae, we get

$$\sum_{i=1}^{n}\sum_{j=1}^{i-1} j \;=\; \frac{1}{2}\left(\frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)}{2}\right) = \frac{n^3}{6} - \frac{n}{6}.$$

(G: 0.5 marks)

Adding the above two expressions, we get

$$\sum_{i=1}^{n}\sum_{j=1}^{i-1}(j-1) + \sum_{i=1}^{n}\sum_{j=1}^{i-1} j \;=\; \frac{n^3}{3} - \frac{n^2}{2} + \frac{n}{6}.$$

Therefore, $a = 1/3$, $b = -1/2$, $c = 1/6$ and $d = 0$.

(H: 0.5 marks)

14. Use Thomas method to obtain the solution for the system

$$\begin{aligned} 2x_1 - x_2 &= -12, \\ -x_1 + 2x_2 - x_3 &= 3, \\ x_2 + 2x_3 - x_4 &= 5, \\ -x_3 + 2x_4 &= -1. \end{aligned}$$

Answer.

Given

$$\alpha_2 = -1,\ \alpha_3 = 1,\ \alpha_4 = -1,$$

$$\beta_1 = 2,\ \beta_2 = 2,\ \beta_3 = 2,\ \beta_4 = 2,$$

$$\gamma_1 = -1,\ \gamma_2 = -1,\ \gamma_3 = -1.$$

The reduced system is

$$x_1 + e_1 x_2 = f_1, \quad e_1 = \frac{\gamma_1}{\beta_1}, \quad f_1 = \frac{b_1}{\beta_1} \;\Rightarrow\; e_1 = -0.5, f_1 = -6.0$$

(A: 0.5 marks)

$$x_2 + e_2 x_3 = f_2, \quad e_2 = \frac{\gamma_2}{\beta_2 - \alpha_2 e_1}, \quad f_2 = \frac{b_2 - \alpha_2 f_1}{\beta_2 - \alpha_2 e_1} \Rightarrow \; e_2 = -0.66667, f_2 = -2.0$$

(B: 0.5 marks)

$$x_3 + e_3 x_4 = f_3, \quad e_3 = \frac{\gamma_3}{\beta_3 - \alpha_3 e_2}, \quad f_3 = \frac{b_3 - \alpha_3 f_2}{\beta_3 - \alpha_3 e_2} \Rightarrow \; e_3 = -0.375, f_3 = 2.625$$

(C: 0.5 marks)

$$(\alpha_4 e_3 - \beta_4) x_4 = \alpha_4 f_3 - b_4 \quad \Rightarrow \; \alpha_4 e_3 - \beta_4 = -1.625, \alpha_4 f_3 - b_4 = -1.625.$$
$$\Rightarrow \; x_4 = 1.$$

(D: 0.5 marks)

Backward substitution gives

$$x_3 + e_3 x_4 = f_3 \implies x_3 - 0.75 \times 1 = 2.25 \implies x_3 = 3.$$

(E: 0.5 marks)

$$x_2 + e_2 x_3 = f_2 \implies x_2 - 0.66667 \times 3 = -2.0 \implies x_2 = 0.$$

and

$$x_1 + e_1 x_2 = f_1 \implies x_1 - 0.5 \times 0 = -6.0 \implies x_1 = -6.$$

(F: 0.5 marks)

Therefore the solution is

$$x_1 = -6.0, x_2 = -0.0, x_3 = 3.0, x_4 = 1.0.$$

The main idea of Thomas algorithm is to write each equation of the system in the form $x_{j+1} + e_{j+1} x_{j+2} = f_{j+1}$ (see "MA214-S23-PART03.pdf" page number 50 in the document). For instance the first equation of the given

system $2x_1 - x_2 = -12$ SHOULD be written in the form $x_1 - 0.5x_2 = -6$. Similarly other equations should be written with the lead term's coefficient as 1 (as shown in steps E and F). No marks are given if the Thomas algorithm is violated even if the solution is obtained correctly but using some other equivalent method. Note that the question is NOT just to solve the system, but to obtain the solution USING THOMAS ALGORITHM.

If all the equations are written in the correct form as per the Thomas algorithm mentioned above and then the back substitution (including 4^{th} equation) process is shown clearly, then the mark for Step G is given.

(G: 1 mark)

15. Let B be an $n \times n$ matrix such that $\rho(B) < 1$. Prove that the sequence $\{\boldsymbol{x}^{(k)}\}$ defined by $\boldsymbol{x}^{(k+1)} = B\boldsymbol{x}^{(k)} + \boldsymbol{c}$ converges to the solution of the system $\boldsymbol{x} = B\boldsymbol{x} + \boldsymbol{c}$ for any initial guess $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$.

[**NOTE:** State (without proof) all the results that you use in the solution clearly and separately AT THE BEGINNING OF THIS SOLUTION.]

Answer.

We need the following two lemmas:

Lemma 1:

Let $B \in M_n(\mathbb{C})$. Then $\rho(B) < 1$ if and only if

$$\lim_{n\to\infty} B^n \boldsymbol{z} = \boldsymbol{0}, \quad \text{for every } \boldsymbol{z} \in \mathbb{C}^n.$$

(A: 0.5 marks)

Lemma 2:

Let $B \in M_n(\mathbb{C})$ with $\rho(B) < 1$. Then $(I - B)^{-1}$ exists and we have

$$(I - B)^{-1} = I + B + B^2 + \dots.$$

(B: 0.5 marks)

Main proof: To prove $\rho(B) < 1 \Rightarrow$ convergence for any $\boldsymbol{x}^{(0)}$.

$$\begin{aligned} \boldsymbol{x}^{(k+1)} &= B\boldsymbol{x}^{(k)} + \boldsymbol{c} \\ &= B\big(B\boldsymbol{x}^{(k-1)} + \boldsymbol{c}\big) + \boldsymbol{c} \\ &= B^2\boldsymbol{x}^{(k-1)} + (B+I)\boldsymbol{c} \end{aligned}$$

Any one of the last two steps is enough.

(C: 0.5 marks)

Continuing in this way, we get

$$\boldsymbol{x}^{(k+1)} = B^{k+1}\boldsymbol{x}^{(0)} + \big(B^k + \ldots + B + I\big)\boldsymbol{c}.$$

(D: 0.5 marks)

Take limit as $k \to \infty$ on both sides to get

$$\lim_{k\to\infty} \boldsymbol{x}^{(k+1)} = \lim_{k\to\infty} B^{k+1}\boldsymbol{x}^{(0)} + \left(\sum_{j=0}^{\infty} B^j\right)\boldsymbol{c}$$

(E: 0.5 marks)

Since $\rho(B) < 1$ we can use Lemma 1 and Lemma 2 to get

$$\lim_{k\to\infty} \boldsymbol{x}^{(k+1)} = (I - B)^{-1}\boldsymbol{c}.$$

Justification is very important. If a student writes this conclusion without quoting the hypothesis and then referring to the above two lemmas, then the marks for Step F are NOT given.

(F (Justification) + G(Conclusion): 0.5 + 0.5 = 1 mark)

Hence, the sequence $\{\boldsymbol{x}^{(k)}\}$ converges and the limit is $\boldsymbol{x} = (I - B)^{-1}\boldsymbol{c}$. By rearranging the terms, we can get $\boldsymbol{x} = B\boldsymbol{x} + \boldsymbol{c}$. This shows that the limit of the sequence $\{\boldsymbol{x}^{(k)}\}$ is the solution of the system $\boldsymbol{x} = B\boldsymbol{x} + \boldsymbol{c}$.

This step is to justify that the limit is indeed the solution of the system. There should be some mathematically correct justification. Otherwise, the marks for this step are NOT given.

(H: 0.5 marks)

16. Consider the matrix

$$A = \begin{pmatrix} 12.25 & 0.125 & 0.42 \\ -1.05 & -14.0 & 0.5 \\ 0.006 & 0.045 & 0.5 \end{pmatrix}.$$

It is known that all the eigenvalues of A are real.

Without calculating the eigenvalues explicitly, show that the eigenvalues of A can be labelled as λ_1, λ_2, and λ_3 satisfying

$$|\lambda_1| > |\lambda_2| > |\lambda_3|.$$

Answer.

Gerschgorin circle theorem applied to A cannot be used to prove the required result. So, we use Gerschgorin circle theorem applied to A^T.

From the **Gerschgorin circle theorem**, we see that all the eigenvalues of A belong to the union of the disks

Student should somewhere mention that the conclusions are made using the GCT. Otherwise, the marks for Step B are not given.

(A: 0.5 marks)

$$\begin{aligned} D_1^T &= \{z \in \mathbb{C} \ : \ |z - 12.25| \leq 1.056\}, \\ D_2^T &= \{z \in \mathbb{C} \ : \ |z + 14.00| \leq 0.17\}, \\ D_3^T &= \{z \in \mathbb{C} \ : \ |z - 0.5| \leq 0.92\}. \end{aligned}$$

If a student writes the disks correctly for A instead of A^T, then these marks are given.

(B + C + D : 0.5+0.5+0.5 = 1.5 marks)

Clearly,

$$\begin{aligned} z \in D_1^T &\implies 11.194 \leq |z| \leq 13.306 \\ z \in D_2^T &\implies 13.83 \leq |z| \leq 14.17 \\ z \in D_3^T &\implies 0 \leq |z| \leq 1.42 \end{aligned}$$

Again if a student writes these estimates for A correctly, then the marks for Step E are given.

(E: 0.5 marks)

Since the eigenvalues are real, we restrict to the intervals

$$I_1 = [11.194, 13.306], \quad I_2 = [-14.17, -13.83], \quad I_3 = [-0.42, 1.42].$$

Again if a student writes these estimates for A correctly, then the marks for Step F are given.

(F: 0.5 marks)

The three intervals $[11.194, 13.306]$, $[13.83, 14.17]$, and $[0, 1.42]$ are disjoint. Note that the intervals have to be written in the absolute sense to conclude the ordering of the absolute values of the eigenvalues.

(G: 0.5 marks)

So, we can say that

$$|\lambda_1| \in [13.83, 14.17], \quad |\lambda_2| \in [11.194, 13.306], \quad |\lambda_3| \in [0, 1.42].$$

(H: 0.5 marks)

Thus, $|\lambda_1| > |\lambda_2| > |\lambda_3|$.

17. Let A and $\tilde{A}$ be $n \times n$ matrices such that A is invertible and

$$\|A - \tilde{A}\| \, \|A^{-1}\| < 1.$$

Let $\boldsymbol{b}$ and $\tilde{\boldsymbol{b}}$ be vectors in $\mathbb{R}^n$ such that $\boldsymbol{b}$ is **non-zero**. Let $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$ be solutions of the systems

$$A\boldsymbol{x} = \boldsymbol{b} \text{ and } \tilde{A}\tilde{\boldsymbol{x}} = \tilde{\boldsymbol{b}},$$

respectively. Prove that

$$\frac{\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|}{\|\boldsymbol{x}\|} \leq \frac{\kappa(A)}{1 - \kappa(A)\frac{\|A-\tilde{A}\|}{\|A\|}} \left(\frac{\|A - \tilde{A}\|}{\|A\|} + \frac{\|\boldsymbol{b} - \tilde{\boldsymbol{b}}\|}{\|\boldsymbol{b}\|} \right)$$

Answer.

From $A\boldsymbol{x} = \boldsymbol{b}$, $\tilde{A}\tilde{\boldsymbol{x}} = \tilde{\boldsymbol{b}}$, we get

$$\boldsymbol{x} - \tilde{\boldsymbol{x}} = -A^{-1}(A - \tilde{A})\tilde{\boldsymbol{x}} + A^{-1}(\boldsymbol{b} - \tilde{\boldsymbol{b}}).$$

(A: 0.5 marks)

On using the properties of norm, and using the subordinate norm for the matrices, we get

$$\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\| \leq \|A^{-1}(A - \tilde{A})\| \, \|\tilde{\boldsymbol{x}}\| + \|A^{-1}\| \, \|\boldsymbol{b} - \tilde{\boldsymbol{b}}\|.$$

(B: 0.5 marks)

Let us divide the last equation with $\|\boldsymbol{x}\|$, and use $\|A\boldsymbol{x}\| = \|\boldsymbol{b}\|$ to get

$$\frac{\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|}{\|\boldsymbol{x}\|} \leq \|A^{-1}(A - \tilde{A})\| \, \frac{\|\tilde{\boldsymbol{x}}\|}{\|\boldsymbol{x}\|} + \|A^{-1}\| \, \frac{\|A\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \frac{\|\boldsymbol{b} - \tilde{\boldsymbol{b}}\|}{\|\boldsymbol{b}\|}.$$

(C: 0.5 marks)

In view of

$$\|\tilde{\boldsymbol{x}}\| = \|\tilde{\boldsymbol{x}} - \boldsymbol{x} + \boldsymbol{x}\| \leq \|\tilde{\boldsymbol{x}} - \boldsymbol{x}\| + \|\boldsymbol{x}\|,$$

the above inequality takes the form

$$\frac{\|\boldsymbol{x}-\tilde{\boldsymbol{x}}\|}{\|\boldsymbol{x}\|} \leq \|A^{-1}(A-\tilde{A})\| \frac{\|\boldsymbol{x}-\tilde{\boldsymbol{x}}\|}{\|\boldsymbol{x}\|} + \|A^{-1}(A-\tilde{A})\| + \|A^{-1}\| \frac{\|A\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \frac{\|\boldsymbol{b}-\tilde{\boldsymbol{b}}\|}{\|\boldsymbol{b}\|}.$$

(D: 0.5 marks)

The above ineqaulity may be re-arranged as

$$\left(1 - \|A^{-1}(A-\tilde{A})\|\right) \frac{\|\boldsymbol{x}-\tilde{\boldsymbol{x}}\|}{\|\boldsymbol{x}\|} \leq \|A^{-1}(A-\tilde{A})\| + \|A^{-1}\| \frac{\|A\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \frac{\|\boldsymbol{b}-\tilde{\boldsymbol{b}}\|}{\|\boldsymbol{b}\|}.$$

Using $\|A\boldsymbol{x}\| \leq \|A\| \, \|\boldsymbol{x}\|$ and $\kappa(A) = \|A^{-1}\| \, \|A\|$, we get

$$\left(1 - \|A^{-1}(A-\tilde{A})\|\right) \frac{\|\boldsymbol{x}-\tilde{\boldsymbol{x}}\|}{\|\boldsymbol{x}\|} \leq \|A^{-1}(A-\tilde{A})\| + \kappa(A) \frac{\|\boldsymbol{b}-\tilde{\boldsymbol{b}}\|}{\|\boldsymbol{b}\|}.$$

(E: 0.5 marks)

Observe that

$$\|A^{-1}(A-\tilde{A})\| \leq \|A^{-1}\| \, \|(A-\tilde{A})\| = \kappa(A) \frac{\|A-\tilde{A}\|}{\|A\|}$$

Observe also that

$$\|A^{-1}(A-\tilde{A})\| \leq \|A^{-1}\| \, \|(A-\tilde{A})\| < 1$$

Thus $1 - \|A^{-1}(A-\tilde{A})\| > 0$.

(F: 0.5 marks)

Therefore, we have

$$\frac{\|\boldsymbol{x}-\tilde{\boldsymbol{x}}\|}{\|\boldsymbol{x}\|} \leq \frac{1}{1 - \|A^{-1}(A-\tilde{A})\|} \left(\|A^{-1}(A-\tilde{A})\| + \kappa(A) \frac{\|\boldsymbol{b}-\tilde{\boldsymbol{b}}\|}{\|\boldsymbol{b}\|} \right).$$

Since $\|A^{-1}(A-\tilde{A})\| \leq \|A^{-1}\|\|A-\tilde{A}\|$, we have

$$\frac{\|\boldsymbol{x}-\tilde{\boldsymbol{x}}\|}{\|\boldsymbol{x}\|} \leq \frac{1}{1-\|A^{-1}(A-\tilde{A})\|}\left(\|A^{-1}\|\,\|(A-\tilde{A})\| + \kappa(A)\frac{\|\boldsymbol{b}-\tilde{\boldsymbol{b}}\|}{\|\boldsymbol{b}\|}\right).$$

(G: 0.5 marks)

Since $\|A^{-1}(A-\tilde{A})\| \leq \|A^{-1}\|\|A-\tilde{A}\|$, we see that

$$\frac{1}{-\|A^{-1}(A-\tilde{A})\|} \leq \frac{1}{-\kappa(A)\frac{\|A-\tilde{A}\|}{\|A\|}}.$$

(H: 0.5 marks)

Using this in the above inequality, we get

$$\frac{\|\boldsymbol{x}-\tilde{\boldsymbol{x}}\|}{\|\boldsymbol{x}\|} \leq \frac{\kappa(A)}{1-\kappa(A)\frac{\|A-\tilde{A}\|}{\|A\|}}\left(\frac{\|A-\tilde{A}\|}{\|A\|} + \frac{\|\boldsymbol{b}-\tilde{\boldsymbol{b}}\|}{\|\boldsymbol{b}\|}\right)$$