MA 214 TSC

Anurag Pendse and Shashwat Chakraborty

14^{th} April 2023

Taylor's Theorem

Taylor's Polynomial

Consider a function f that is n-times continuously differentiable at a given point a. The **Taylor's polynomial** of degree n for f is defined by

$$T_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!} (x-a)^k$$

Taylor's Theorem

Let f be $(n+1)$-times differentiable on an open interval containing the points a and x. There exists a number $\xi \in (a, x)$ (or (x, a) depending on which is bigger) such that

$$f(x) = T_n(x) + \frac{f^{(n+1)}(\xi)}{(n+1)!} (x-a)^{n+1}$$

Taylor's Theorem

Mean Value Theorem for Integration

(i) If f is continuous on $[a, b]$, then there exists a number c in $[a, b]$ such that

$$\int_a^b f(x)dx = f(c)(b-a)$$

(ii) Let $f(x)$ and $g(x)$ be continuous on the interval $[a, b]$, and let $g(x) \geq 0$ for all $x \in \mathbb{R}$. There exists a number $c \in [a, b]$ such that

$$\int_a^b f(x)g(x)dx = f(c)\int_a^b g(x)dx$$

Problems

1. If n is a positive integer, show that

$$\int_{\sqrt{n\pi}}^{\sqrt{(n+1)\pi}} \sin(t^2)dt = \frac{(-1)^n}{c}$$

where $\sqrt{n\pi} \leq c \leq \sqrt{(n+1)\pi}$

Order of Convergence

Big Oh and Little oh

Let $\{a_n\}$ and $\{b_n\}$ be sequences of real numbers.

(i) We write $a_n = O(b_n)$, if there exists a real number C and a natural number N such that

$$|a_n| \leq C|b_n|, \quad \text{for all } n \geq N$$

(ii) We write $a_n = o(b_n)$, if for every $\epsilon > 0$ $\exists N \in \mathbb{N}$ such that

$$|a_n| \leq \epsilon|b_n|, \quad \text{for all } n \geq N$$

Arithmetic Error Analysis

Floating Point Representation

Any number can be represented in base $\beta \in \mathbb{N}$ $(\beta \geq 2)$ as

$$(-1)^s \times (0.d_1 d_2 d_3 \cdots)_\beta \times \beta^e$$

(i) s is called the sign

(ii) $(0.d_1 d_2 d_3 \cdots)_\beta$ is called the mantissa

(iii) e is called the exponent

$n-$digit floating point representation

An $n-$digit floating point number in base β is of the form

$$(-1)^s \times (0.d_1 d_2 d_3 \cdots d_n)_\beta \times \beta^e$$

Arithmetic Error Analysis

Arithmetic using $n-$digit rounding and chopping

Let $\odot$ denote any of the arithmetic operations $+, -, \div$, and $\times$. To calculate $x \odot y$ using $n-$digit rounding/chopping arithmetic, follow the step-by-step procedure given below:

(i) Write the $n-$digit rounding/chopping approximation of x and y: $\text{fl}(x)$ and $\text{fl}(y)$

(ii) Perform $\text{fl}(x) \odot \text{fl}(y)$ exactly

(iii) Get the $n-$digit approximation of the result $\text{fl}\left(\text{fl}(x) \odot \text{fl}(y)\right)$

Significance

Significant $\beta-$digits

Let x be a real number $\neq 0$. Let x_A be an approximation to x. We say that x_A approximates x to r significant $\beta-$digits if r is the largest non-negative integer such that

$$\frac{|x - x_A|}{|x|} \leq 0.5 \times \beta^{-r+1}$$

Propagation of Error

Condition Number

The condition number of a function f at a point c is given by

$$\kappa = \left| \frac{f'(c)}{f(c)} c \right|$$

A process is said to be well-conditioned if the condition number is sufficiently small. Otherwise, it's called ill-conditioned.

Problems

1. Consider the function

$$f(x) = \sqrt{x+1} - \sqrt{x}, \quad \text{for all } x \in [0, \infty)$$

   Check the stability of the computational process.

2. Let x, y, and z be real numbers whose floating-point approximations in a computing device coincide with the numbers themselves. Show that the relative error in calculating $x \times (y + z)$ equals $\epsilon_1 + \epsilon_2 - \epsilon_1\epsilon_2$, where $\epsilon_1 = E_r(\text{fl}(y + z))$ and $\epsilon_2 = E_r(\text{fl}(x \times \text{fl}(y + z)))$.

Gaussian Elimination with Partial Pivoting

Consider the following system of three linear equations:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3 \end{aligned}$$

Procedure:

(i) Define $s_1 = \max\{|a_{11}|, |a_{21}|, |a_{31}|\}$. Let $s_1 = |a_{k1}|$. Interchange the first and the k^{th} equations.

$$\begin{aligned} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + a_{13}^{(1)}x_3 &= b_1^{(1)} \\ a_{21}^{(1)}x_1 + a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 &= b_2^{(1)} \\ a_{31}^{(1)}x_1 + a_{32}^{(1)}x_2 + a_{33}^{(1)}x_3 &= b_3^{(1)} \end{aligned}$$

Then perform the first step of Naive Gaussian Elimination

(ii) The new system looks like

$$\begin{aligned} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + a_{13}^{(1)}x_3 &= b_1^{(1)} \\ 0 + a_{22}^{(2)}x_2 + a_{23}^{(2)}x_3 &= b_2^{(2)} \\ 0 + a_{32}^{(2)}x_2 + a_{33}^{(2)}x_3 &= b_3^{(2)} \end{aligned}$$

Define $s_2 = \max\{|a_{22}^{(2)}|, |a_{32}^{(2)}|\}$. Let $s_2 = |a_{l2}^{(2)}|$. Interchange the second and the l^{th} equations.Then perform the second step of the Naive Gaussian Elimination.

(iii) Use backward substitution to obtain the solution to the system of linear equations

Thomas Method

Tri-diagonal Systems

A tri-diagonal system is of the following form:

$$
\begin{aligned}
\beta_1 x_1 + \gamma_1 x_2 + 0x_3 + 0x_4 \cdots + 0x_{n-1} + 0x_n &= b_1 \\
\alpha_2 x_1 + \beta_2 x_2 + \gamma_2 x_3 + 0x_4 + \cdots + 0x_{n-1} + 0x_n &= b_2 \\
0x_1 + \alpha_3 x_2 + \beta_3 x_3 + \gamma_3 x_4 + \cdots + 0x_{n-1} + 0x_n &= b_3 \\
&\vdots \\
0x_1 + 0x_2 + \cdots + \alpha_{n-1} x_{n-2} + \beta_{n-1} x_{n-1} + \gamma_{n-1} x_n &= b_{n-1} \\
0x_1 + 0x_2 + 0x_3 + \cdots + 0x_{n-2} + \alpha_n x_{n-1} + \beta_n x_n &= b_n
\end{aligned}
$$

Thomas Method

(i) Write the first equation as

$$x_1 + e_1 x_2 = f_1, \quad e_1 = \frac{\gamma_1}{\beta_1}, f_1 = \frac{b_1}{\beta_1}$$

Eliminate x_1 from the second equation.

(ii) Rewrite the resulting equation as

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

(iii) Proceeding similarly, we get the j^{th} equation as

$$x_j + e_j x_{j+1} = f_j$$

Eliminate x_j from the $(j+1)^{\text{th}}$ equation.

Thomas Method

(iv) The modified $(j+1)^{\text{th}}$ equation looks like

$$x_{j+1} + e_{j+1} x_{j+2} = f_{j+1}$$

where

$$e_{j+1} = \frac{\gamma_{j+1}}{\beta_{j+1} - \alpha_{j+1} e_j}$$
$$f_{j+1} = \frac{b_{j+1} - \alpha_{j+1} f_j}{\beta_{j+1} - \alpha_{j+1} e_j}$$

(v) We finally obtain the reduced n^{th} equation as

$$(\alpha_n e_{n-1} - \beta_n)\, x_n = \alpha_n f_{n-1} - b_n$$

(vi) Use backward substitution to get $x_n, x_{n-1}, \cdots, x_2, x_1$

LU Factorization

Doolittle Factorization

If $n \geq 2$ and A be an $n \times n$ invertible matrix such that all of its $n-1$ leading principal minors are non-zero. Then A has an $LU-$decomposition where the lower triangular matrix has a unit diagonal.

Cholesky Factorization

If A is an $n \times n$ symmetric and positive definite matrix, then A has a unique factorization

$$A = LL^T$$

where L is a lower triangular matrix with positive diagonal elements.

Positive Definite Matrix

A symmetric matrix A is said to be positive definite if

$$\mathbf{x}^T A\mathbf{x} > 0$$

for all non-zero vectors $\mathbf{x}$. The following statements are all equivalent:

(i) A is positive definite.
(ii) All the principal minors of A are positive.
(iii) All the eigenvalues of the matrix A are positive.

Matrix Norms

Vector Norm

A vector norm on $\mathbb{R}^n$ is a function $\|\cdots\| : \mathbb{R}^n \to [0, \infty)$ with the following properties:

(i) $\|\mathbf{x}\| \geq 0$ for all vectors $\mathbf{x} \in \mathbb{R}^n$

(ii) $\|\mathbf{x}\| = 0$ iff $\mathbf{x} = \mathbf{0}$

(iii) $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|$ for all $\mathbf{x} \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$

(iv) $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ (Triangle Inequality)

Matrix Norms

Three important vector norms;

(i) Euclidean Norm (l_2-norm):

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2}$$

(ii) Maximum Norm (l_∞-norm):

$$\|\mathbf{x}\|_\infty = \max\{|x_1|, |x_2|, |x_3|, \cdots, |x_n|\}$$

(iii) l_1-norm:

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|$$

Matrix Norms

Matrix Subordinate Norms

Let $\|\cdot\|$ be a vector norm on $\mathbb{R}^n$ and let $A \in M_n(\mathbb{R})$. The subordinate matrix norm is defined as

$$\|A\| := \sup\{\|A\mathbf{x}\| : x \in \mathbb{R}^n, \|x\| = 1\}$$

Properties of Subordinate Norms

Let $\|\cdot\|$ be a subordinate norm. Then

(i) $\|A\mathbf{x}\| \leq \|A\|\|\mathbf{x}\|$ for all $A \in M_n(\mathbb{R})$ and $\mathbf{x} \in \mathbb{R}^n$

(ii) $\|\mathbb{I}\| = 1$

(iii) $\|AB\| \leq \|A\|\|B\|$ for all $A, B \in M_n(\mathbb{R})$

Matrix Norms

Important Subordinate Norms

(i) Matrix norm subordinate to the maximum norm: (max row sum)

$$\|A\|_\infty = \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|$$

(ii) Matrix norm subordinate to the l_1-norm: (max column sum)

$$\|A\|_1 = \max_{1\le i\le n} \sum_{i=1}^{n} |a_{ij}|$$

(iii) Matrix norm subordinate to the Euclidean norm:

$$\|A\|_2 = \sqrt{\max_{1\le i\le n} |\lambda_i|}$$

where $\lambda_1, \lambda_2, \cdots, \lambda_n$ are eigenvalues of the matrix $A^T A$.

Matrix Norms

Theorem

Let A be an invertible matrix $n \times n$ matrix. Let $\mathbf{x}$ and $\tilde{\mathbf{x}}$ be the solutions of the systems

$$A\mathbf{x} = \mathbf{b} \text{ and } A\tilde{\mathbf{x}} = \tilde{\mathbf{b}}$$

Then

$$\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \leq \|A\| \, \|A^{-1}\| \frac{\|\mathbf{b} - \tilde{\mathbf{b}}\|}{\|\mathbf{b}\|}$$

Condition Number

The condition number of a matrix A is defined as

$$\kappa(A) := \|A\|\|A^{-1}\|$$

If the condition number is small, then the relative error in the solution will be small whenever the relative error in the RHS vector is small.

Problems

1. Find the Doolittle factorization of the matrix

$$A = \begin{bmatrix} 60 & 30 & 20 \\ 30 & 20 & 15 \\ 20 & 15 & 12 \end{bmatrix}$$

2. Let $\|\cdot\|$ be a matrix norm subordinate to a fixed vector norm. Then show that if D is a matrix such that $\|D\| < 1$, then the matrix $I - D$ is non-singular.

Jacobi Method

We wish to solve the system $A\mathbf{x} = \mathbf{b}$.

(i) Write $A = D - C$, where D is diagonal.

(ii) $D\mathbf{x} = C\mathbf{x} + b$

(iii) Start with an initial guess $\mathbf{x}^{(0)}$ to obtain a better approximation $\mathbf{x}^{(1)}$.

(iv) Repeat the procedure to obtain the subsequent iterates

$$D\mathbf{x}^{(i+1)} = C\mathbf{x}^{(i)} + \mathbf{b}$$

If D is invertible, then

$$\mathbf{x}^{(i+1)} = B\mathbf{x}^{(i)} + \mathbf{c}$$

where $B = D^{-1}C$ and $\mathbf{c} = D^{-1}\mathbf{b}$

Diagonally Dominant Matrix

A matrix A is diagonally dominant if it satisfies the inequality

$$\sum_{j=1, j\neq i}^{n} |a_{ij}| < |a_{ii}|, \ \ i = 1, 2, 3, \cdots, n$$

Convergence theorem for Jacobi method

If the coefficient matrix is diagonally dominant, then the Jacobi method converges to the exact solution of $A\mathbf{x} = \mathbf{b}$

Gauss-Siedel Method

Gauss-Siedel Iterative Sequence

We wish to solve the linear system $A\mathbf{x} = \mathbf{b}$. To demonstrate the method, we take the 3×3 case.

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3 \end{aligned}$$

Iteration sequence: (for $k = 0, 1, 2, \cdots$)

$$\begin{aligned} x_1^{(k+1)} &= \frac{1}{a_{11}}\left(b_1 - a_{12}x_2^{(k)} - a_{13}x_3^{(k)}\right) \\ x_2^{(k+1)} &= \frac{1}{a_{22}}\left(b_2 - a_{21}x_1^{(k+1)} - a_{23}x_3^{(k)}\right) \\ x_3^{(k+1)} &= \frac{1}{a_{33}}\left(b_3 - a_{31}x_1^{(k+1)} - a_{32}x_3^{(k+1)}\right) \end{aligned}$$

Gauss-Siedel Method

Convergence Theorem for GS method

If the coefficient matrix is diagonally dominant, then the Gauss-Siedel method converges to the exact solution.

Spectral Radius

Let $A \in M_n\mathbb{C}$ and let $\lambda_i \in \mathbb{C},\ i = 1, 2, \cdots, n$, be the eigenvalues of A. The spectral radius of A is defined as

$$\rho(A) = \max_{j=1,2,\cdots,n} |\lambda_j|$$

For each $A \in M_n\mathbb{C}$ and each $\epsilon > 0$, there exist a subordinate norm such that

$$\rho(A) \leq \|A\| < \rho(A) + \epsilon$$

Necessary and Sufficient Condition for Convergence

For any $\mathbf{x^{(0)}} \in \mathbb{R}$, the sequence of iterates $\{\mathbf{x^{(k)}}\}$ converges to the solution of $\mathbf{x} = B\mathbf{x} + \mathbf{c}$ iff

$$\rho(B) < 1$$

Eigenvalues and Eigenvectors

Dominant Eigenvalue of a Matrix

An eigenvalue λ of an $n \times n$ matrix A is said to be a dominant eigenvalue if

$$|\lambda| = \max\{|z| : z \text{ is an eigenvalue of } A\}$$

Hypotheses for Power Method

(H1) The matrix A has a unique, simple dominant eigenvalue (say λ_1).

(H2) The matrix A is diagonalizable.

(H3) The initial guess $x^{(0)}$ is chosen such that

$$\mathbf{x}^(0) = \sum_{i=1}^{n} c_i \mathbf{v}_i$$

where $\{v_i\}$ is the eigenbasis of A, and c_j's are some scalars with $c_1 \neq 0$

Eigenvalues and Eigenvectors

Power Method

(i) Choose an initial guess $\mathbf{x}^{(0)}$ that satisfies (H3).

(ii) Generate the sequences $\{\mu_k\}$ and $\{x^{(k)}\}$ using

$$\mu_{k+1} = y_i^{(k+1)}$$

$$\mathbf{x}^{(k+1)} = \frac{\mathbf{y}^{(k+1)}}{\mu_{k+1}}$$

where $y^{(k+1)} = Ax^{(k)}$

Convergence theorem for Power Method

If the hypotheses (H1)-(H3) are met, then

(i) the sequence $\{\mu_k\}$ converges to the dominant eigenvalue (λ_M)

(ii) a subsequence of $\{\mathbf{x}^{(k)}\}$ converges to an eigenvector corresponding to λ_M

Convergence theorem for Power Method

If the hypotheses (H1)-(H3) are met, then

(i) the sequence $\{\mu_k\}$ converges to the dominant eigenvalue (λ_M)

(ii) a subsequence of $\{\mathbf{x}^{(k)}\}$ converges to an eigenvector corresponding to λ_M

If we make an additional hypothesis as given below

(H4) The eigenvector $\mathbf{v}_1$ corresponding to the maximum eigenvalue (say λ_1) has a single maximum component

then the sequence $\{\mathbf{x}_k\}$ converges to an eigenvector corresponding to the dominant eigenvalue.

Gershgorin's Circle Theorem

Let A be an $n \times n$ matrix. For each $k = 1, 2, \cdots, n$, define ρ_k as

$$\rho_k = \sum_{j \neq k} |a_{kj}|$$

Let D_k denote the disk centered at a_{kk} with radius ρ_k.

(i) Each eigenvalue lies in one of the disks.

(ii) Suppose that among the n disks, a collection of m disks is disjoint from the rest. Then exactly m eigenvalues lie in the union of the m disks, and $n - m$ lie in the union of the remaining disks.

Interpolation

Lagrange Interpolating Polynomial

1. $l_i(x) = \prod_{k \neq i} \frac{x - x_k}{x_i - x_k}$
2. $p_n(x) = \sum_{i=0}^{n} f(x_i) l_i(x)$

Newton's Form of Interpolating Polynomial

$$p_n(x) = A_0 + A_1(x - x_0) + \cdots + A_n \prod_{i=0}^{n-1}(x - x_i)$$

Divided Differences

1. $p_n(x) = f[x_0] + f[x_0, x_1](x - x_0) + \cdots + f[x_0, \ldots, x_n] \prod_{i=0}^{n-1} (x - x_i)$
2. The divided differences are invariant under permutation of the arguments

Divided Differences

$$f[x_0, \ldots, x_n] = \frac{f[x_1, \ldots, x_n] - f[x_0, \ldots, x_{n-1}]}{x_n - x_0}$$

Interpolation

Mathematical Error

$$ME_n(x) = \frac{f^{n+1}(\xi_x)}{(n+1)!} \prod_{i=0}^{n} (x - x_i) \qquad \xi_x \in (a, b)$$

Divided Differences

$$f[x_0, \ldots, x_n, x] = \frac{f^{(n+1)}(\xi_x)}{(n+1)!}$$

Interpolation

Arithmetic Error

$$|AE_n| \leq |\epsilon| \sum_{i=0}^{n} ||l_i(x)||_{\infty,I}$$

Total Error

$$|TE_n| \leq ||f - p_n||_{\infty,I} + |\epsilon| \sum_{i=0}^{n} ||l_i(x)||_{\infty,I}$$

Theorem (Faber)

Let the sequence of nodes $a < x_0^{(n)} < \cdots < x_n^{(n)} < b$ be given. Then there exists a continuous function on $[a, b]$ such that the polynomials p_n that interpolate f have the property that $||p_n - f||_{\infty,[a,b]}$ does not tend to zero as $n \to \infty$

Theorem

Let a function f which is continuous on $[a, b]$ be given. Then there exists a sequence of nodes $a < x_0^{(n)} < \cdots < x_n^{(n)} < b$ such that the polynomials p_n interpolating f on these nodes have the property that $||p_n - f||_{\infty,[a,b]}$ tends to zero as $n \to \infty$

Piecewise Interpolation

In this method, we interpolate the function on a subset of the nodes while covering all the nodes.
Assume $n+1$ nodes. The degree of each interpolating polynomial will be less than n. Say the degree of the required polynomial is k. Then we will choose the first $k+1$ nodes and interpolate f on them. Then do the same on the next $k+1$, with the first in this set being the last one from the previous set.

Problems

1. Carry out piecewise polynomial interpolation using polynomials of degree 2 for $f(x) = \cos(x)$ at the points $x = 0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}, 2\pi$

Nonlinear Equations

Closed Domain Methods

We need to know an interval where at least one root of the nonlinear equation lies. We then restrict that interval with successive iterations.

Open Domain Methods

We can start with an arbitrary guess and refine that with successive iterations.

Bisection method

1. Define x_1 to be the midpoint of the interval $[a_0, b_0]$ with $f(a_0)f(b_0) < 0$.
2. Calculate $f(x_1)$. If it is 0, then we have found the root. Else, move ahead
3. The new interval has its endpoints as a_1 and b_1 where one of a_1 or b_1 is x_1, and the other is one of the original endpoints where the sign of the function at that endpoint has an opposite sign to that at x_1
4. Repeat steps 1-3

Regula-Falsi Method

1. $x_1 = \frac{a_0 f(b_0) - b_0 f(a_0)}{f(b_0) - f(a_0)}$
2. The new interval has its endpoints as a_1 and b_1 where one of a_1 or b_1 is x_1, and the other is one of the original endpoints where the sign of the function at that endpoint has an opposite sign to that at x_1
3. Repeat steps 1-2

Secant Method

1. $x_{n+1} = \frac{x_{n-1}f(x_n) - x_n f(x_{n-1})}{f(x_n) - f(x_{n-1})}$

Convergence of Secant Method

$$\lim_{n\to\infty} \frac{|x_{n+1} - r|}{|x_n - r|^\alpha} = \frac{|f''(r)|}{|2f'(r)|} \qquad \alpha = \frac{1+\sqrt{5}}{2}$$

Newton-Raphson Method

1. $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$

Convergence of Newton-Raphson Method

$$\lim_{n\to\infty} \frac{|x_{n+1} - r|}{|x_n - r|^2} = \frac{|f''(r)|}{|2f'(r)|}$$

Newton-Raphson Method on Convex functions

If f is a convex, monotonically increasing function with at least one real root, then it has a unique root and the Newton-Raphson Method converges to it for all initial guesses x_0.

Fixed Point Iteration Method

1. We want to solve an equation of the form $g(x) = x$.
2. $x_{n+1} = g(x_n)$
3. If the sequence $\{x_i\}$ converges, then we say that the Fixed Point Iteration method converges.

Convergence of the Fixed Point Iteration Method

If g is a continuously differentiable self-map that is also a contraction map, then

1. g has a unique root in the interval.
2. The fixed point iteration converges to that root irrespective of the initial guess
3. $|x_n - r| \leq \lambda^n |x_0 - r| \leq \frac{\lambda^n}{1-\lambda} |x_1 - x_0|$
4. $\lim_{n \to \infty} \frac{x_{n+1} - r}{x_n - r} = g'(r)$

Problems

1. Find the value of $\sqrt{5}$
2. Is the order of convergence of the Newton-Raphson method for $f(x) = (x-1)^2$ two?

Numerical Differentiation

Forward Difference Formula

$$D_h^+ f = \frac{f(x+h) - f(x)}{h}$$

$$ME = -\frac{h}{2} f''(\eta)$$

Backward Difference Formula

$$D_h^- f = \frac{f(x) - f(x-h)}{h}$$

Central Difference Formula

$$D_h^0 f = \frac{f(x+h) - f(x-h)}{2h}$$

$$ME = -\frac{h^2}{6} f'''(\eta)$$

Mathematical Error

Let $f \in C^{n+2}[a,b]$. Let $x_0, x_1, \ldots, x_n$ be nodes in $[a,b]$. Let p_n interpolate f at these points. Let x be a point in $[a,b]$ that is not equal to any of the nodes. Then

$$f'(x) - p_n'(x) = w_n(x)\frac{f^{(n+2)}(\eta_x)}{(n+2)!} + w_n'(x)\frac{f^{(n+1)}(\xi_x)}{(n+1)!}$$

Where $w_n(x) = \prod_{i=0}^{n}(x - x_i)$, ξ_x and η_x lie between the max and min amongst the nodes and are dependent on x.

Method of Undetermined Coefficients

$$f^{(k)}(x) = w_0 f(x_0) + \cdots + w_n f(x_n)$$

We find the coefficients w_i by imposing the condition that the formula is exact for all polynomials of order less than or equal to n

Newton-Cotes Formula

Assume equally spaced nodes (quadrature points). Denote the quadrature weights by $w_i = \int_a^b l_i dx = I(l_i)$ where l_i is the i^{th} Lagrange interpolating polynomial at the nodes. Then

$$I(f) = w_0 f(x_0) + w_1 f(x_1) + \dots$$

Numerical Integration

Rectangle Rule

$$I(f) = (b-a)f(x_0)$$

$$I_R(f) = (b-a)f(a)$$

Midpoint Rule

$$I_M(f) = (b-a)f(\frac{a+b}{2})$$

Error

$$ME_R(f) = \frac{f'(\eta)(b-a)^2}{2}$$

Numerical Integration

Trapezoidal Rule

$$I_T(f) = (b-a)\left(\frac{f(a)+f(b)}{2}\right)$$

Error

$$ME_T(f) = -\frac{f''(\eta)(b-a)^3}{12}$$

Composite Trapezoidal Rule

$$I_T^n(f) = h\left(\frac{1}{2}f(x_0) + f(x_1) + \cdots + f(x_{n-1}) + \frac{1}{2}f(x_n)\right)$$

Numerical Integration

Simpson's Rule

$$I_s(f) = \frac{(b-a)}{6}(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b))$$

Error

$$ME_S(f) = -\frac{f^{(4)}(\eta)(b-a)^5}{2880}$$

Composite Simpson's Rule

$$I_S^n(f) = \frac{h}{3}\left(f(x_0) + f(x_{2n}) + 2\sum_{i=1}^{n-1} f(x_{2i}) + 4\sum_{i=0}^{n-1} f(x_{2i+1})\right)$$

Method of Undetermined Coefficients

$$I(f) = \sum_{i=0}^{n} w_i f(x_i)$$

Find the coefficients by imposing the condition that the formula is exact for polynomials of degree less than or equal to n.

Gaussian Rules

$$\int_{-1}^{1} f(x)dx = \sum_{i=0}^{n} w_i f(x_i)$$

Numerical Integration

n=0

$$I_{G0}(f) = 2f(0)$$

n=1

$$I_{G1}(f) = f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right)$$

Problems

1. Find the approximate formula for the second derivative of a function
2. Let $f : [a, b] \to \mathbb{R}$ be a twice continuously differentiable function. Derive an expression for the mathematical error involved in approximating the integral $\int_a^b f(x)dx$ using the mid-point rule.

Fundamental Theorem

A continuous function y defned on an interval I containing the point x_0 is a solution of the initial value problem $y' = f(x, y), y(x_0) = y_0$ if and only if y satisfies the integral equation

$$y = y_0 + \int_{x_0}^{x} f(x, y)dx \qquad \forall x \in I$$

Cauchy-Lipschitz-Picard's Existence and Uniqueness Theorem

Let $D \subseteq \mathbb{R}^2$ be a domain and $I \subset \mathbb{R}$ be an interval. Let $f : D \to \mathbb{R}$ be a continuous function. Let $(x_0, y_0) \in D$ be a point such that the rectangle $\mathbb{R}$ defined by $R = \{x : |x - x_0| \leq a\} \times \{y : |y - y_0| \leq b\}$ is contained in D. Let f be Lipschitz continuous with respect to the variable y on R, i.e., there exists a $K > 0$ such that

$$|f(x, y_1) - f(x, y_1)| \leq K|y_1 - y_2| \ \forall (x, y_1), (x, y_2) \in R.$$

Then the initial value problem has at least one solution on the interval $|x - x_0| \leq \delta$ where $\delta = \min(a, \frac{b}{M})$. Moreover, the initial value problem has exactly one solution on this interval.

Existence and Uniqueness Theorem

Let $D \subseteq \mathbb{R}^2$ be a domain and $I \subseteq \mathbb{R}$ be an interval. Let $f : D \to \mathbb{R}$ be a continuous function. Let $(x_0, y_0) \in D$ be a point such that the rectangle R defined by

$$R = \{x : |x - x_0| \leq a\} \times \{y : |y - y_0| \leq b\}$$

is contained in D. If the partial derivative $\frac{\partial f}{\partial y}$ is also continuous in D, then there exists a unique solution $y = y(x)$ of the initial value problem defined on the interval $|x - x_0| \leq \delta$ where $\delta = \min(a, \frac{b}{M})$

Numerical ODEs

Forward Euler's Method

$$y_{j+1} = y_j + hf(x_j, y_j)$$

Backward Euler's Method

$$y_{j-1} = y_j - hf(x_j, y_j)$$

Error in Euler's Method

$$T_{j+1} = \frac{h^2}{2} y''(\xi_j) \qquad \xi_j \in (x_j, x_{j+1})$$

$$\text{Propgated error} = y(x_{j)} - y_j + h(f(x_j, y(x_j)) - f(x_j, y_j))$$

$$ME(y_{j+1}) = \left[1 + h\frac{\partial f}{\partial y}(x_j, \eta_j)\right] ME(y_j) + \frac{h^2}{2} y''(\xi_j) \qquad \eta_j \in (y(x_j), y_j)$$

Bound for ME

If $|\frac{\partial f}{\partial y}| < L$ and $y''(x) < Y$ in the interval concerned, then

$$ME(y_j) \leq \frac{hY}{2L}(e^{(x_n - x_0)L} - 1) + e^{(x_n - x_0)L}|y(x_0) - y_0|$$

Bound on Total Error

If $|\frac{\partial f}{\partial y}| < L$ and $y''(x) < Y$ in the interval concerned, $y_j = \tilde{y}_j + \epsilon_j$, $\epsilon = \max(|\epsilon_j|)$,then

$$TE(y_j) \leq \frac{1}{L}\left(\frac{hY}{2} + \frac{\epsilon}{h}\right)(e^{(x_n - x_0)L} - 1) + e^{(x_n - x_0)L}|\epsilon_0|$$

Euler's Midpoint Method

$$y_{j+1} = y_{j-1} + 2hf(x_j, y_j)$$

Euler's Trapezoidal Method

$$y_{j+1} = y_j + \frac{h}{2}(f(x_j, y_j) + f(x_{j+1}, y_{j+1}))$$

Runge-Kutta Method of order 2

$$y_{j+1} = y_j + \frac{h}{2}(k_1 + k_2)$$

$$k_1 = f(x_j, y_j)$$

$$k_2 = f(x_{j+1}, y_j + hk_1)$$

Runge-Kutta Method of order 4

$$y_{j+1} = y_j + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4)$$

$$k_1 = f(x_j, y_j)$$

$$k_2 = f\left(x_j + \frac{h}{2}, y_j + \frac{h}{2}k_1\right)$$

$$k_3 = f\left(x_j + \frac{h}{2}, y_j + \frac{h}{2}k_2]\right)$$

$$k_4 = f(x_j + h, y_j + hk_3)$$

Doubts?

Contact

Thank You and All the Best!

Contact info:

Anurag - 8975971044

Shashwat - 9340262841