Numerical Analysis

gouziwu

March 29, 2019

Contents

1 Chap1 Mathematical Preliminaries

1.1 1.2 Roundoff Errors and Computer Arithmetic

Truncation Error : the error involved in using a truncated, or finite, summation to approximate the sum of an infinite series

Roundoff Error: the error produced when performing real number calculations. It occurs because the arithmetic performed in a machine involves numbers with only a finite number of digits.

Suppose $y = 0.d_1d_2\dots d_kd_{k+1}d_{k+2}\dots\times10^n$, then

$$fl(y) = \begin{cases} 0.d_1d_2\dots d_k\times10^n & \text{chopping} \\ chop(y+5\times10^{n-(k+1)}) = 0.\delta_1\delta_2\dots\delta_k\times10^n & \text{Rounding} \end{cases}$$

Definition 1.1. *If $p*$ is an approximation to p, the absolute error is $|p-p*|$, and the relative error is $\frac{|p-p*|}{|p|}$, provided that $p\neq0$*

Definition 1.2. *The number $p*$ is said to approximate p to t significant digits if t is the largest nonnegative integer for which $\frac{|p-p*|}{|p|} < 5\times10^{-t}$*

chopping $|\frac{y-fl(y)}{y}| = |\frac{0.d_1d_2\dots d_kd_{k+1}\dots\times10^n - 0.d_1d_2\dots d_k\times10^n}{0.d_1d_2\dots d_kd_{k+1}\times10^n}| = |\frac{0.d_{k+1}\dots}{0.d_1d_2\dots}|\times10^{-k}\leqslant \frac{1}{0.1}\times10^{-k} = 10^{-k+1}$

rounding $|\frac{y-fl(y)}{y}| \leqslant \frac{0.5}{0.1}\times10^{-k} = 0.5\times10^{-k+1}$

Finite digit arithmetic

- $x\oplus y = fl(fl(x)+fl(y))$
- $x\otimes y = fl(fl(x)\times fl(y))$
- $x\ominus y = fl(fl(x)-fl(y))$
- $x \oplus\!\!\!\!\div\, y = fl(fl(x)\div fl(y))$

1.2 1.3 ALgorithms and Convergence

An algorithm that satisfies that small changes in the initial data produce correspondingly small changes in the final results is called **stable**; otherwise it is **unstable**. An algorithm is called **conditionally stable** if it is stable only for certain choices of initial data.

Suppose that E >0 denotes an initial error and En represents the magnitude of an error after n subsequent operations. If $E_n \approx CnE_0$, where C is a constant independent of n, then the growth of error is said to be **linear**. If $E_n\approx C^nE_0$, for some C >1, then the growth of error is called **exponential**

Suppose $\{\beta_n\}_{n=1}^{\infty}$, $\lim\limits_{n\to\infty}\beta_n = 0$, $\{\alpha_n\}_{n=1}^{\infty}$, $\lim\limits_{n\to\infty}\alpha_n = \alpha$. If a positive constant K exists with $|\alpha_n-\alpha|\leqslant K|\beta_n|$ for large n, then $\{\alpha_n\}_{n=1}^{\infty}$ converges to with **rate, or order, of convergence** $O(\beta_n)$

Suppose $\lim\limits_{h\to0}G(h) = 0$, $\lim\limits_{h\to0}F(h) = L$ and $|F(h)-L|\leqslant K|G(h)|$ for sufficiently small h, then we write $F(h) = L+O(G(h))$

2 Chap2 Solutions of equations in one variable

2.1 2.1 Bisection method

Theorem 2.1. *Intermediate Value Theorem If $f \in C[a,b]$, $K \in (f(a), f(b))$, then there exists a number $p \in (a,b)$ for which $f(p) = K$*

Theorem 2.2. *Suppose that $f \in C[a,b]$ and $f(a) \cdot f(b) < 0$. The bisection method generates a sequence $\{p_n\}, n = 0, 1, \dots$ approximating a zero p of f with*

$$|p_n - p| \leqslant \frac{b-a}{2^n}, \quad \text{when } n \geqslant 1$$

2.2 2.2 Fixed-Point Iteration

$f(x) = 0 \xleftrightarrow{\text{equivalent}} x = f(x) + x = g(x)$

Theorem 2.3. *Fixed-Point Theorem Let $g \in C[a,b]$ be s.t. $g(x) \in [a,b]$ for all $x \in [a,b]$. Suppose that g' exists on (a,b) and that a constant $0 < k < 1$ exists with $|g'(x)| \leqslant k$ for all $x \in (a,b)$ (hence g' can't converge to 1). Then for any number p_0 in $[a,b]$, the sequence defined by $p_n = g(p_{n-1}), n \geqslant 1$ converges to the unique point p in $[a,b]$*

Corollary 2.1. $|p_n - p| \leqslant \frac{1}{1-k}|p_{n+1} - p_n|$ *and* $|p_n - p| \leqslant \frac{k^n}{1-k}|p_1 - p_0|$

2.3 2.3 Newton's method

Linearize a nonlinear function using **Taylor's expansion**

Let $p_0 \in [a,b]$ be an approximation to p s.t. $f'(p_0) \neq 0$, hence $f(x) = f(p_0) + f'(p_0)(x - p_0) + \frac{f''(\xi_x)}{2!}(x - p_0)^2$, then $0 = f(p) \approx f(p_0) + f'(p_0)(p - p0) \to p \approx p_0 - \frac{f(p_0)}{f'(p_0)}$ $p_n = p_{n-1} - \frac{f(p_{n-1})}{f'(p_{n-1})}, \quad \text{for} n \geqslant 1$

Theorem 2.4. *Let $f \in C^2[a,b]$. If $p \in [a,b]$ is s.t. $f(p) = 0, f'(p) \neq 0$, then there exists a $\delta > 0$ s.t. Newton's method generates a sequence $\{p_n\}, n \in \mathbb{N}\backslash\{0\}$ converging to p for any initial approximation $p \in [p - \delta, p + \delta]$.*

2.4 2.4 Error analysis for iterative methods

Definition 2.1. *Suppose $\{p_n\}(n = 0, 1, \dots)$ is a sequence that converges to p with $p_n \neq p$ for all n. If positive constants α and λ exist with*

$$\lim_{n \to \infty} \frac{|p_{n+1} - p|}{|p_n - p|^\alpha} = \lambda$$

then $\{p_n\}(n = 0, 1, \dots)$ *converges to* p *of order* α, *with asymptotic error constant* λ

Theorem 2.5. *Let* p *be a fixed point of* $g(x)$. *If there exists some constant* $\alpha \geqslant 2$ *s.t.* $g \in C^{\alpha}[p-\delta, p+\delta]$, $g'(p) = \cdots = g^{\alpha-1}(p) = 0$ *and* $g^{\alpha}(p) \neq 0$. *Then the iterations with* $p_n = g(p_{n-1})$, $n \geqslant 1$ *is of order* α

$$p_{n+1} = g(p_n) = g(p) + g'(p)(p_n - p) + \cdots + \frac{g^{\alpha}(\xi_n)}{\alpha!}(p_n - p)^{\alpha}$$

Theorem 2.6. *Let* $g \in C[a, b]$ *be s.t.* $g(x) \in [a, b]$ *for all* $x \in [a, b]$. *Suppose in addition that* g' *is continuous on* (a, b) *and a positive constant* $k < 1$ *exists with*

$$|g'(x)| \leqslant k, \quad \text{for all } x \in (a, b)$$

If $g'(p) \neq 0$, *then for any number* $p_0 \neq p$ *in* $[a, b]$, *the sequence*

$$p_n = g(p_{n-1}), \quad \text{for } n \geqslant 1$$

converges only linearly to the unique fixed point in $[a, b]$

Proof.

$$\begin{aligned}
\lim_{n\to\infty} \frac{|p_{n+1} - p|}{|p_n - p|} &= \lim_{n\to\infty} \frac{|g(p_n) - p|}{|p_n - p|} \\
&= \lim_{n\to\infty} \frac{|g'(\xi)(p_n - p)|}{|p_n - p|} \\
&= |g'(p)|
\end{aligned}$$

□

Theorem 2.7. *Let* p *be a solution of the equation* $x = g(x)$. *Suppose that* $g'(p) = 0$ *and g" is continuous with* $|g''(x)| < M$ *on an open interval* I *containing* p. *Then there exists a* $\delta > 0$ *s.t. for* $p_0 \in [p-\delta, p+\delta]$, *the sequence defined by* $p_n = g(p_{n-1})$, *when* $n \geqslant 1$ *converges at least quadratically to* p. *Moreover, for sufficiently large values of* n,

$$|p_{n+1} - p| < \frac{M}{2}|p_n - p|^2$$

Proof. Choose $k \in (0,1), \delta > 0$ s.t. $[p-\delta, p+\delta] \subseteq I$ and $|g'(x)| < k$ and g'' is continuous.

$$g(x) = g(p) + g'(p)(x-p) + \frac{g''(\xi)}{2}(x-p)^2$$

Hence $g(x) = p + \frac{g''(\xi)}{2}(x-p)^2$. $p_{n+1} = g(p_n) = p + \frac{g''(\xi_n)}{2}(p_n - p)^2$. Thus $p_{n+1} - p = \frac{g''(\xi_n)}{2}(p_n-p)^2$. We get

$$\lim_{n\to\infty} \frac{|p_{n+1}-p|}{|p_n - p|^2} = \frac{g''(p)}{2}$$

□

Definition 2.2. *A solution p of $f(x) = 0$ is a zero of multiplicity m of f if for $x \neq p$, $f(x) = (x-p)^m q(x)$ where $\lim_{x\to p} q(x) \neq 0$*

Theorem 2.8. *The function $f \in C^m[a,b]$ has a zero of multiplicity m at p in (a,b) if and only if*

$$0 = f(p) = f'(p) = \cdots = f^{(m-1)}(p), \quad \textit{but } f^{(m)}(p) \neq 0$$

To handle the problem of multiple roots of a function f is to define $\mu(x) = \frac{f(x)}{f'(x)}$.

If p is a zero of f of multiplicity m with $f(x) = (x-p)^m q(x)$, then

$$\begin{aligned}
\mu(x) &= \frac{(x-p)^m q(x)}{m(x-p)^{m-1}q(x) + (x-p)^m q'(x)} \\
&= (x-p)\frac{q(x)}{mq(x) + (x-p)q'(x)}
\end{aligned}$$

And $q(x) \neq 0$.

Now Newton's method:

$$\begin{aligned}
g(x) &= x - \frac{\mu(x)}{\mu'(x)} \\
&= x - \frac{f(x)/f'(x)}{(f'(x)^2 - f(x)f''(x))/f'(x)^2} \\
&= x - \frac{f(x)f'(x)}{f'(x)^2 - f(x)f''(x)}
\end{aligned}$$

3 Chap6 Direct Methods for Solving Linear Systems

3.1 6.1 Linear Systems of Equations

Gaussian elimination with backward substitution

3.2 6.2 Pivoting Strategies

Problem: small pivot element may cause trouble

Paritial Pivoting: Determine the smallest pk s.t. $|a_{pk}^{(k)}| = \max_{k \leqslant j \leqslant n} |a_{ik}^{(k)}|$ and interchange the pth and the kth rows

Scaled Partial Pivoting:

1. Define a scale factor s_i for each row as $s_i = \max_{1 \leqslant j \leqslant n} |a_{ij}|$
2. Determine the smallest $p \geqslant k$ s.t. $\frac{|a_{pk}^{(k)}|}{s_p} = \max_{k \leqslant i \leqslant n} \frac{|a_{ik}^{(k)}|}{s_i}$ and interchange the pth and the kth rows

Complete Pivoting: Search all the entries a_{ij} to find the entry with the largest magnitude

3.3 6.5 Matrix Factorization

$m_{ik} = a_{ik}/a_{kk}$

$$L_k = \begin{pmatrix} 1 & & & & \\ & \ddots & & \text{\Large 0} & \\ & & 1 & & \\ & & -m_{k+1,k} & & \\ & & \vdots & \ddots & \\ & & -m_{n,k} & & 1 \end{pmatrix}$$

Hence

$$L_1^{-1} L_2^{-1} \dots L_{n-1}^{-1} = \begin{pmatrix} 1 & & & \text{\Large 0} \\ & 1 & & \\ & & \ddots & \\ \text{\Large $m_{i,j}$} & & & 1 \end{pmatrix}$$

$$U = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ & a_{22} & \dots & a_{2n} \\ & & \dots & \vdots \\ & & & a_{nn} \end{pmatrix}$$

$A = LU$

3.4 6.6 Special Types of Matrices

Strictly Diagonally Dominant Matrix. $|a_{ii}| > \sum_{\substack{j=1, \\ j \neq i}}^{n} |a_{ij}| \quad \text{for each } i = 1, \dots, n$

Theorem 3.1. *A strictly diagonally dominant matrix A is nonsingular. Moreover, Gaussian elimination can be performed without row or column interchanges, and the computations will be stable w.r.t. the growth of roundoff errors*

Choleski's Method for Positive Definite Matrix:

Definition 3.1. *A matrix A is positive definite if ti's symmetric and if* $\mathbf{x}^T \mathbf{A} \mathbf{x} > 0$ *for every n-dimensional vector* $\mathbf{x} \neq 0$

Lemma 3.1. *A is positive definite*

1. A^{-1} *is positive definite as well, and* $a_{ii} > 0$
2. $\sum |a_{ij}| \leqslant \max |a_{kk}|$; $(a_{ij})^2 < a_{ii} a_{jj}$ *for each i j*
3. *Each of /A's leading principal submatrices* A_k*/ has a positive determinant*

$$U = \begin{pmatrix} & u_{ij} & \end{pmatrix} = \begin{pmatrix} u_{11} & & \\ & \ddots & \\ & & u_{nn} \end{pmatrix} \begin{pmatrix} 1 & & u_{ij}/u_{ii} \\ & 1 & \\ & & 1 \end{pmatrix} = D\tilde{U}$$

A is symmetric, hence

$$L = \tilde{U}^t, A = LDL^t$$

Let

$$D^{1/2} = \begin{pmatrix} \sqrt{u_{11}} & & \\ & \ddots & \\ & & \sqrt{u_{nn}} \end{pmatrix}, \tilde{L} = LD^{1/2/}, A = \tilde{L}\tilde{L}^t$$

Crout Reduction for tridiagonal Linear System

$$\begin{pmatrix} b_1 & c_1 & & & \\ a_2 & b_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & a_{n-1} & b_{n-1} & c_{n-1} \\ & & & a_n & b_n \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_{n-1} \\ f_n \end{pmatrix}$$

$$A = \begin{pmatrix} \alpha_1 & & & \\ \gamma_2 & \ddots & & \\ & \ddots & \ddots & \\ & & \gamma_n & \alpha_n \end{pmatrix} \begin{pmatrix} 1 & \beta_1 & & \\ & \ddots & \ddots & \\ & & \ddots & \beta_{n-1} \\ & & & 1 \end{pmatrix}$$

4 Chap7 Iterative techiniques in Matrix algebra

4.1 7.1 Norms of vectors and matrices

Definition 4.1. *A vector norm on R^n is a function $||\cdot|| : \mathbb{R}^n \to \mathbb{R}$ with following properties for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n, \alpha \in C$*

1. $||\mathbf{x}|| \leqslant 0$; $||\mathbf{x}|| = 0 \Longleftrightarrow \mathbf{x} = \mathbf{0}$
2. $||\alpha\mathbf{x}|| = |\alpha| \cdot ||\mathbf{x}||$
3. $||\mathbf{x} + \mathbf{y}|| \leqslant ||\mathbf{x}|| + ||\mathbf{y}||$

$||\mathbf{x}||_1 = \sum_{i=1}^{n} |x_i|$. $||\mathbf{x}_p|| = (\sum_{i=1}^{n} |x_i|^p)^{1/p}$

Definition 4.2. *A sequence $\{\mathbf{x}^{(k)}\}_{k=1}^{\infty}$ of vectors in R^n converge to $\mathbf{x}$ w.r.t the norm $||\cdot||$ if given any $\epsilon > 0$ there exists an integer $N(\epsilon)$ s.t. $||\mathbf{x}^{(k)} - \mathbf{x}|| < \epsilon$ for all $k \geqslant N(\epsilon)$*

Theorem 4.1. *The sequence of vectors $\{\mathbf{x}^{(k)}\}$ converges to $\mathbf{x} \in R^n$ w.r.t. $||\cdot||$ if and only if $\lim_{k\to\infty} \mathbf{x}_i^{(k)} = x_i$ for each $i = 1, 2, \ldots, n$*

Definition 4.3. *If there exist positive constants* C_1, C_2 *s.t.* $C_1||\mathbf{x}||_B \leqslant ||\mathbf{x}||_A \leqslant C_2||\mathbf{x}|_B|$. *Then* $||\cdot||_A, ||\cdot||_B$ *are equivalent*

Theorem 4.2. *All the vector norm in* R^n *are equivalent*

Definition 4.4. *A matrix norm on the set of* $n \times n$*:*

1. $||\mathbf{A}|| \geqslant 0; ||\mathbf{A}|| = 0 \iff \mathbf{A} = \mathbf{0}$
2. $||\alpha\mathbf{A}|| = |\alpha| \cdot ||\mathbf{A}||$
3. $||\mathbf{A} + \mathbf{B}|| \leqslant ||\mathbf{A}|| + ||\mathbf{B}||$
4. $||\mathbf{AB}|| \leqslant ||\mathbf{A}|| \cdot ||\mathbf{B}||$

Frobenius Norm: $||\mathbf{A}||_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}|^2}$

Natural Norm: $||\mathbf{A}||_p = \max_{\mathbf{x}\neq\mathbf{0}} \frac{||\mathbf{Ax}||_p}{||\mathbf{x}||} = \max_{\mathbf{z}\neq\mathbf{0}} ||\mathbf{A}\frac{\mathbf{z}}{||\mathbf{z}||}|| = \max_{||\mathbf{x}||_p=1} ||\mathbf{Ax}||_p$