Numerical Analysis

gouziwu

April 28, 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 $y = 0.d_1 d_2 \dots d_k d_{k+1} d_{k+2} \dots \times 10^n$, then

$$fl(y) = \begin{cases} 0.d_1 d_2 \dots d_k \times 10^n & \text{chopping} \\ chop(y + 5 \times 10^{n-(k+1)}) = 0.\delta_1 \delta_2 \dots \delta_k \times 10^n & \text{Rounding} \end{cases}$$

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 $p \neq 0$*

Definition 1.2. *The number $p*$ is said to approximate p to t significant digits if t is the largest nonnegative integer for which $\frac{|p-p*|}{|p|} < 5 \times 10^{-t}$*

chopping $|\frac{y-fl(y)}{y}| = |\frac{0.d_1 d_2 \dots d_k d_{k+1} \dots \times 10^n - 0.d_1 d_2 \dots d_k \times 10^n}{0.d_1 d_2 \dots d_k d_{k+1} \times 10^n}| = |\frac{0.d_{k+1} \dots}{0.d_1 d_2 \dots}| \times 10^{-k} \leqslant \frac{1}{0.1} \times 10^{-k} = 10^{-k+1}$

rounding $|\frac{y-fl(y)}{y}| \leqslant \frac{0.5}{0.1} \times 10^{-k} = 0.5 \times 10^{-k+1}$

Finite digit arithmetic

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

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 $E_n \approx C^n E_0$, 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 $\lim\limits_{h\to 0} G(h) = 0, \lim\limits_{h\to 0} F(h) = L$ 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 $\{p_n\}, n = 0, 1, \ldots$ approximating a zero p of f with*

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

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.

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

$\square$

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

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

$\square$

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*

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

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 Chap3 Interpolation and polynomial approximation

3.1 3.1 Interpolation and the Lagrange polynomial

$P_n(x) = \sum_{i=0}^{n} L_{n,i}(x)y_i$. Find $L_{n,i}(x)$ for $i = 0, \ldots, n$ s.t. $L_{n,j}(x_j) = \delta_{ij}$. δ_{ij} Kronecker delta. Each $L_{n,i}$ has n roots $x_0, \ldots, \hat{x_i}, \ldots, x_n$. $L_{n,j}(x) = C_i(x-x_0)\ldots(x \hat{-} x_i)\ldots(x-x_n) = C_i \prod_{\substack{j\neq i \\ j=0}}^{n}(x-x_j)$. $L_{n,j}(x_i) = 1 \to C_i = \prod_{j\neq i} \frac{1}{x_i - x_j}$. Hence $L_{n,i}(x) = \prod_{\substack{j\neq i \\ j=0}}^{n} \frac{x-x_j}{x_i-x_j}$

Theorem 3.1. *If $x_0, x_1, \ldots, x_n$ are n+1 distinct numbers and f is a function whose values are given at these numbers, then the n-th Lagrange interpolating polynomial is unique*

Analyze the remainder. Suppose $a \leqslant x_0 < x_1 < \cdots < x_n \leqslant b$ and $f \in C^{n+1}[a,b]$. Consider $R_n(x) = f(x) - P_n(x)$. $R_n(x)$ has at least n+1 roots => $R_n(x) = K(x)\prod_{i=1}^{n}(x-x_i)$. For any $x \neq x_i$. Define $g(t) = R_n(t) - K(x)\prod_{i=0}^{n}(t-x_i)$. $g(x)$ has n+2 distinct roots $x_0 \ldots x_n x$. Hence $g^{(n+1)}(\xi_x) = 0, \xi_x \in (a,b)$. $f^{(n+1)}(\xi_x) - Pn^{(n+1)}(\xi_x) - K(x)(n+1)! = R_n^{(n+1)}(\xi_x) - K(x)(n+1)!$. Thus $R_n(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!}\prod_{i=0}^{n}(x-x_i)$.

Definition 3.1. *Let f be a function defined at $x_0, \ldots, x_n$ and suppose $m_1, \ldots, m_k$ are k distinct integers with $0 \leqslant m_i \leqslant n$ for each i. The Lagrange polynomial that agrees with $f(x)$ at the k points $x_{m_1}, \ldots, x_{m_k}$ denoted by $P_{m_1,;m_k}(x)$*

Theorem 3.2. *Let f be defined at $x_0, \ldots, x_k$ and let x_i and x_j be two distinct numbers in this set. Then*

$$P(x) = \frac{(x-x_j)P_{0,1,\ldots,j-1,j+1,\ldots,k(x)} - (x-x_i)P_{0,\ldots,i-1,i+1,\ldots,k(x)}}{x_i - x_j}$$

describes the k-th Lagrange polynomial that interpolates f at the k+1 points $x_0, \ldots, x_k$

Neville's Method

$$\begin{array}{lllll} x_0 & P_0 & & & \\ x_1 & P_1 & P_{0,1} & & \\ x_2 & P_2 & P_{1,2} & P_{0,1,2} & \\ x_3 & P_3 & P_{2,3} & P_{1,2,3} & P_{0,1,2,3} \end{array}$$

3.2 3.2 Divied differences

$f[x_i, x_j] = \frac{f(x_i)-f(x_j)}{x_i-x_j}(i \neq j, x_i \neq x_j)$. $f[x_i, x_j, x_k] = \frac{f[x_i,x_j]-f[x_j,x_k]}{x_i-x_k}$.

3.3 Additional Newton Interpolation

3.3.1 Simple idea

Given $x_0, \ldots, x_n$

1. Fitting x_0 first: $f(x) \approx f_0, f_0 = f(x_0)$
2. Add one more point x_1, $f_1 = f(x_1)$

$$f(x) \approx f_0 + \alpha_1(x-x_0), \alpha_1 = \frac{f_1 - f_0}{x_1 - x_0}$$

3. More points $f(x) \approx f_0 + \alpha_1(x - x_0) + \alpha_2(x - x_0)(x - x_1)$

The pattern and coefficients. $f(x) = \sum_{i=0}^{n} \alpha_i \prod_{j=0}^{j<i}(x - x_j) = \sum_{i=0}^{n} \alpha_i N^{(i)}(x)$

$$\begin{pmatrix} f_0 \\ f_1 \\ \vdots \\ f_n \end{pmatrix} = \begin{pmatrix} N^{(0)}(x_0) & N^{(1)}(x_0) & \dots & N^{(n)}(x_0) \\ N^{(0)}(x_1) & N^{(1)}(x_1) & \dots & N^{(n)}(x_1) \\ \vdots & \vdots & \ddots & \vdots \\ N^{(0)}(x_n) & N^{(1)}(x_n) & \dots & N^{(n)}(x_n) \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix}$$

$N^{(i)}(x_k) = \begin{cases} 0 & k < i \\ \prod_{j=0}^{j<i}(x_k - x_j) & k \geqslant i \end{cases}$ with $N^{(0)}(x) = 1$. Newton interpolation matrix is lower triangular. Lagrange matrix is identity.

3.3.2 Basis transformation

$$\begin{pmatrix} 1 \\ (x - x_0) \\ (x - x_0)(x - x_1) \\ \vdots \end{pmatrix} = (?) \begin{pmatrix} 1 \\ x \\ x^2 \\ \vdots \end{pmatrix}$$

Hence $(\Phi_B)^T = (T_A^B)^T (\Phi_A)^T$. $\Phi_B = \Phi_A T_A^B$

$$\begin{aligned} (\Phi_A)(\alpha_A) = (f) &= (\Phi_B)(\alpha_B) \\ &= (\Phi_A)(T_A^B)(\alpha_B) \\ &\Rightarrow \\ (\alpha_A) &= (T_A^B)(\alpha_B) \\ (\alpha_B) &= (T_A^B)^{-1}(\alpha_A) \\ &= (T_B^A)(\alpha_A) \end{aligned}$$

4 Chap6 Direct Methods for Solving Linear Systems

4.1 6.1 Linear Systems of Equations

Gaussian elimination with backward substitution

4.2 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

4.3 6.5 Matrix Factorization

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

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

Hence

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

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

$A = LU$

4.4 6.6 Special Types of Matrices

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

Theorem 4.1. *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 4.1. *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 4.1. *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

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

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

5 Chap7 Iterative techiniques in Matrix algebra

5.1 7.1 Norms of vectors and matrices

Definition 5.1. *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 5.2. *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 5.1. *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 5.3. *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 5.2. *All the vector norm in R^n are equivalent*

Definition 5.4. *A matrix norm on the set of $n \times n$:*

1. $||\mathbf{A}|| \geqslant 0; ||\mathbf{A}|| = 0 \Longleftrightarrow \mathbf{A} = \mathbf{0}$
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: $||\mathbf{A}||_p = \max_{\mathbf{x}\neq\mathbf{0}} \frac{||\mathbf{Ax}||_p}{||\mathbf{x}||_p} = \max_{\mathbf{z}\neq\mathbf{0}} ||\mathbf{A}\frac{\mathbf{z}}{||\mathbf{z}||}|| = \max_{||\mathbf{x}||_p=1} ||\mathbf{Ax}||_p$

$||\mathbf{A}||_\infty = \max_{1\leqslant i\leqslant n}\sum_{j=1}^{n}|a_{ij}|$, $||\mathbf{A}||_1 = \max_{1\leqslant j\leqslant n}\sum_{i=1}^{n}|a_{ij}|$, $||\mathbf{A}||_2 = \sqrt{\lambda_{\max}(\mathbf{A}^T\mathbf{A})}$

5.2 7.2 Eigenvalues and Eigenvectors

spectral radius.

Definition 5.5. *The spectral radius $\rho(A)$ of a matrix A is defined as $\rho(A) = \max|\lambda|$ where λ is an eigenvalue of A*

Theorem 5.3. *If A is an $n \times n$ matrix, then $\rho(A) \leqslant ||A||$ for any natural norm*

Proof. $|\lambda| \cdot ||\boldsymbol{x}|| = ||\lambda\boldsymbol{x}|| = ||A\boldsymbol{x}|| \leqslant ||A|| \cdot ||\boldsymbol{x}||$ □

Definition 5.6. *We call an $n \times n$ matrix A convergent if for all $i, j = 1, \dots, n$* $\lim_{k\to\infty}(A^k)_{ij} = 0$

5.3 7.3 Iterative techniques for solving linear systems

Jacobi iterative method.

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \cdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n \end{cases} \Longrightarrow \begin{cases} x_1 = \frac{1}{a_{11}}(-a_{12}x_2 - \cdots - a_{1n}x_n + b_1) \\ x_2 = \frac{1}{a_{22}}(-a_{21}x_1 - \cdots - a_{2n}x_n + b_2) \\ \cdots \\ x_1 = \frac{1}{a_{nn}}(-a_{n2}x_1 - \cdots - a_{nn-1}x_{n-1} + b_n) \end{cases}$$

In matrix form,

$$A = \begin{pmatrix} D & -U & -U \\ -L & D & -U \\ -L & -L & D \end{pmatrix}$$

$$
\begin{aligned}
A\boldsymbol{x} = \boldsymbol{b} &\Leftrightarrow (D - L - U)\boldsymbol{x} = \boldsymbol{b} \\
&\Leftrightarrow D\boldsymbol{x} = (L + U)\boldsymbol{x} + \boldsymbol{b} \\
&\Leftrightarrow \boldsymbol{x} = \underbrace{D^{-1}(L + U)}_{T_j}\boldsymbol{x} + \underbrace{D^{-1}}_{\boldsymbol{c}_j}\boldsymbol{b}
\end{aligned}
$$

. T_j is Jacobi iterative matrix. $\boldsymbol{x}^{(k)} = T_j\boldsymbol{x}^{(k-1)} + \boldsymbol{c}_j$
Gauss-Seidel iterative method

$$
\begin{aligned}
&\boldsymbol{x}^{(k)} = D^{-1}(L\boldsymbol{x}^{(k)} + U\boldsymbol{x}^{(k-1)}) + D^{-1}\boldsymbol{b} \\
\Leftrightarrow &(D - L)\boldsymbol{x}^{(k)} = U\boldsymbol{x}^{(k-1)} + \boldsymbol{b} \\
\Leftrightarrow &\boldsymbol{x}^{(k)} = \underbrace{(D - L)^{-1}U\boldsymbol{x}^{(\boldsymbol{k-1})}}_{T_g} + \underbrace{(D - L)^{-1}\boldsymbol{b}}_{\boldsymbol{c}_g}
\end{aligned}
$$

convergence of iterative methods

Theorem 5.4. *the following are equivalent:*

1. *A is a convergent matrix*
2. $\lim_{n\to\infty} ||A^n|| = 0$ *for some natural norm*
3. $\lim_{n\to\infty} ||A^n|| = 0$ *for all natural norms*
4. $\rho(A) < 1$
5. $\lim_{n\to\infty} A^n\boldsymbol{x} = \boldsymbol{0}$ *for every* $\boldsymbol{x}$

$\boldsymbol{e}^{(k)} = \boldsymbol{x}^{(k)} - \boldsymbol{x}^* = (T\boldsymbol{x}^{(k-1)} + \boldsymbol{c}) - (T\boldsymbol{x}^* + \boldsymbol{c}) = T(\boldsymbol{x}^{(k-1)} - \boldsymbol{x}^*) = T\boldsymbol{e}^{(k-1)} \Rightarrow \boldsymbol{e}^{(k)} = T^k\boldsymbol{e}^{(0)}$. $||\boldsymbol{e}^{(k)} \leqslant ||T|| \cdot ||\boldsymbol{e}^{(k-1)}|| \leqslant \cdots \leqslant ||T||^k \cdot ||ble^{(0)}||$

Theorem 5.5. *For any $\boldsymbol{x}^{(0)} \in R^n$, the sequence $\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty}$ defined by $\boldsymbol{x}^{(k)} = T\boldsymbol{x}^{(k-1)} + \boldsymbol{c}$ for each k, converges to the unique solution of $\boldsymbol{x} = T\boldsymbol{x} + \boldsymbol{c}$ if and only if $\rho(T) < 1$*

$\rho(T) < 1 \Longrightarrow (I - T)^{-1} =$
$displaystyle \sum_{j=0}^{\infty} T^j$

Theorem 5.6. *If $\|T\| < 1$ for any natural matrix norm and $\boldsymbol{c}$ is a given vector, then the sequence $\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty}$ defined by $\boldsymbol{x}^{(k)} = T\boldsymbol{x}^{(k-1)} + \boldsymbol{c}$ converges for any $\boldsymbol{x}^{(0)} \in R^n$ to a vector $\boldsymbol{x}$. And the following error bounds hold*

1. $\left\|\boldsymbol{x}-\boldsymbol{x}^{(k)}\right\| \leqslant \|T\|^k\left\|\boldsymbol{x}-\boldsymbol{x}^{(0)}\right\|$

2. $\left\|\boldsymbol{x}-\boldsymbol{x}^{(k)}\right\| \leqslant \frac{\|T\|^k}{1-\|T\|}\left\|\boldsymbol{x}^{(1)}-\boldsymbol{x}^{(0)}\right\|$

Theorem 5.7. *If A is a strictly diagonally dominant, then for any choice of $\boldsymbol{x}^{(0)}$, both the Jacobi and Gauss-Seidel methods give sequences $\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty}$ that converges to the unique solution*

relaxation methods. $x_i^{(k)} = \frac{1}{a_{ii}}(b_i - \sum\limits_{j=1}^{i-1} a_{ij}x_i^{(k)} - \sum\limits_{j=i+1}^{n} a_{ij}x_j^{(k-1)}) = x_i^{(k-1)} + \dfrac{r_i^{(k)}}{a_{ii}}$ and relaxation method is $x_i^{(k)} = x_i^{(k-1)} + \omega\frac{r_i^{(k)}}{a_{ii}}$

Theorem 5.8. *(kahan) If $a_{ii} \neq 0$ for each i. Then $\rho(T_\omega) \geqslant |\omega - 1|$.*

This implies the SOR method can converge only if $0 < \omega < 2$

Theorem 5.9. *(Ostrowski-Reich) If A is positive definite and $0 < \omega < 2$, the SOR converges*

Theorem 5.10. *If A is positive definite and tridiagonal, then $\rho(T_g) = (\rho(T_j))^2 < 1$, and the optimal choice of ω for the SOR method is $\omega = \frac{2}{1+\sqrt{1-(\rho(T_j))^2}}$. With this choice of ω, we have $\rho(T_\omega) = \omega - 1$*

5.4 7.4 Error bounds and iterative refinement

Assume that A is accurate and $\boldsymbol{b}$ has the error $\delta\boldsymbol{b}$, then $\boldsymbol{A}(\boldsymbol{x}+\delta\boldsymbol{x}) = \boldsymbol{b}+\delta\boldsymbol{b}$

Theorem 5.11. *Suppose $\tilde{\boldsymbol{x}}$ is an approximation to the solution of $\boldsymbol{Ax} = \boldsymbol{b}$ A is nonsingular matrix. Then for any natural norm,*

$$||\boldsymbol{x} - \tilde{\boldsymbol{x}}|| \leqslant ||\boldsymbol{r}|| \cdot ||A^{-1}||$$

and if $\boldsymbol{x} \neq \boldsymbol{0}, \boldsymbol{b} \neq \boldsymbol{0}$,

$$\frac{||\delta\boldsymbol{x}||}{||\boldsymbol{x}||} \leqslant ||\boldsymbol{A}|| \cdot ||\boldsymbol{A}^{-1}|| \cdot \frac{||\delta\boldsymbol{b}||}{||\boldsymbol{b}||}$$

Proof. $\boldsymbol{r} = \boldsymbol{b} - \boldsymbol{A}\tilde{\boldsymbol{x}} = A\boldsymbol{x} - A\tilde{\boldsymbol{x}}$ and A is nonsingular. Hence $\boldsymbol{x} - \tilde{\boldsymbol{x}} = A^{-1}\boldsymbol{r}$. Since $\frac{||A^{-1}\boldsymbol{r}||}{||\boldsymbol{r}||} \leqslant ||A^{-1}||$, $||\boldsymbol{x} - \tilde{\boldsymbol{x}}|| = ||A^{-1}\boldsymbol{x}|| \leqslant ||A^{-1}1|| \cdot ||\boldsymbol{r}||$. Also $||\boldsymbol{b}|| \leqslant ||A|| \cdot ||\boldsymbol{x}||$. So $1/||\boldsymbol{x}|| \leqslant ||A||/||\boldsymbol{b}||$ □

Theorem 5.12. *If a matrix B satisfies $||B|| < 1$ for some natural norm, then*

1. *$I \pm B$ is nonsingular*
2. *$||(I \pm B)^{-1}|| \leqslant \frac{1}{1-||B||}$*

Assume $\boldsymbol{b}$ is accurate, A has the error δA, then $(A+\delta A)(\boldsymbol{x}+\delta\boldsymbol{x}) = \boldsymbol{b}$. Hence $\frac{||\delta\boldsymbol{x}||}{||\boldsymbol{x}||} \leqslant \frac{||A^{-1}||\cdot||\delta A||}{1-||A^{-1}||\cdot||\delta A||} = \frac{||A||\cdot||A^{-1}||}{1-||A||\cdot||A^{-1}||\cdot\frac{||\delta A||}{||A||}}$

condition number K(A) is $||A|| \cdot ||A^{-1}||$

Theorem 5.13. *Suppose A is nonsingular and $||\delta A|| \leqslant \frac{1}{||A^{-1}||}$. The solution $\boldsymbol{x}+\delta\boldsymbol{x}$ to $(A+\delta A)(\boldsymbol{x}+\delta\boldsymbol{x})$ approximates the solution $\boldsymbol{x}$ of $A\boldsymbol{x} = \boldsymbol{b}$ with the error estimate*

$$\frac{||\delta\boldsymbol{x}||}{||\boldsymbol{x}||} \leqslant \frac{K(A)}{1-K(A)||\delta A||/||A||}\left(\frac{||\delta A||}{||A||} + \frac{||\delta\boldsymbol{b}||}{||\boldsymbol{b}||}\right)$$

note:

1. If A is symmetric, then $K(A)_2 = \frac{\max|\lambda|}{\min|\lambda|}$
2. $K(A)_p \geqslant 1$ for all natural norm
3. $K(\alpha A)_= K(A)$ for any $\alpha \in R$
4. $K(A)_2 = 1$ if A is orthogonal
5. $K(RA)_2 = K(AR)_2 = K(A)_2$ for all orthogonal matrix R_

iterative refinement:

Theorem 5.14. *Suppose $\boldsymbol{x}^*$ is an approximation to the solution of $A\boldsymbol{x} = \boldsymbol{b}$, A is nonsingular matrix and $\boldsymbol{r} = \boldsymbol{b} - A\boldsymbol{x}$. Then for any natural norm, $||\boldsymbol{x} - \boldsymbol{x}^* \leqslant ||\boldsymbol{r}|| \cdot ||A^{-1}||$, and if $\boldsymbol{x}, \boldsymbol{b} \neq \boldsymbol{0}$*

$$\frac{||\boldsymbol{x}-\boldsymbol{x}^*||}{||\boldsymbol{x}||} \leqslant K(A)\frac{||\boldsymbol{r}||}{||\boldsymbol{b}||}$$

refinement

1. $A\boldsymbol{x} = \boldsymbol{b}$ => approximation $\boldsymbol{x}_1$
2. $\boldsymbol{r}_1 = \boldsymbol{b} - A\boldsymbol{x}_1$
3. $A\boldsymbol{d}_1 = \boldsymbol{r}_1$ => $\boldsymbol{d}_1$
4. $\boldsymbol{x}_2 = \boldsymbol{x}_1 + \boldsymbol{d}_1$

6 chap9 Approximating Eigenvalues

6.1 9.3 the power method

the original method Assumptions: A is an $n \times n$ matrix with eigenvalues satisfying $|\lambda_1| > |\lambda_2| \geqslant \cdots \geqslant |\lambda_n| \geqslant 0$

$$
\begin{aligned}
&\boldsymbol{x}^{(0)} = \sum_{j=1}^{n} \beta_j \boldsymbol{v}_j, \quad \beta_1 \neq 0 \\
&\boldsymbol{x}^{(1)} = A\boldsymbol{x}^{(0)} = \sum_{j=1}^{n} \beta_j \lambda_j \boldsymbol{v}_j \\
&\boldsymbol{x}^{(2)} = A\boldsymbol{x}^{(1)} = \sum_{j=1}^{n} \beta_j \lambda_j^2 \boldsymbol{v}_j \\
&\cdots \\
&\boldsymbol{x}^{(k)} \approx \lambda_1^k \beta_1 \boldsymbol{v}_1, \quad \lambda_1 \approx \frac{\boldsymbol{x}_i^{(k)}}{\boldsymbol{x}_i^{(k-1)}}
\end{aligned}
$$

Normalization. Suppose $||\boldsymbol{x}||_\infty = 1$. Let $||\boldsymbol{x}^{(k)}||_\infty = |x_{p_k}^{(k)}|$.Then $\boldsymbol{u}^{(k-1)} = \frac{\boldsymbol{x}^{(k-1)}}{|x_{p_{k-1}}^{(k-1)}|}$ and $\boldsymbol{x}^{(k)} = A\boldsymbol{u}^{(k-1)}$. Then $\boldsymbol{u}^{(k)} = \frac{\boldsymbol{x}^{(k)}}{|x_{p_k}^{(k)}|} \to \boldsymbol{v}_1$. $\lambda_1 \approx \frac{\boldsymbol{x}_i^{(k)}}{\boldsymbol{u}_i^{(k-1)}} = \boldsymbol{x}_{p_{k-1}}^{(k)}$

Note:

1. the method works for **multiple** eigenvalues $\lambda_1 = \lambda_2 = \cdots = \lambda_r$
2. the method fails to converge if $\lambda_1 = -\lambda_2$
3. Aitken's Δ^2 can be used

Rate of convergence. $\boldsymbol{x}^{(k)} = A\boldsymbol{x}^{(k-1)} = \lambda_1^k \sum_{j=1}^{n} \beta_j (\frac{\lambda_j}{\lambda_1})^k \boldsymbol{v}_j$. Make $|\lambda_2/\lambda_1|$ as small as possible. Assume $\lambda_1 > \lambda_2 \geqslant \cdots \geqslant \lambda_n, |\lambda_2| > |\lambda_n|$. Let $B = A - pI$, then $|\lambda I - A| = |\lambda I - (B + pI)| = |(\lambda - p)I - B|$. Hence $\lambda_A - p = \lambda_B$. Since $\frac{|\lambda_2 - p|}{|\lambda_1 - p|} < \frac{|\lambda_2|}{|\lambda_1|}$. The iteration is fast

Inverse power method. If A has $|\lambda_1| \geqslant |\lambda_2| \geqslant \cdots > |\lambda_n|$, then A^{-1} has $|\frac{1}{\lambda_n}| > |\frac{1}{\lambda_{n-1}}| \geqslant \cdots \geqslant |\frac{1}{\lambda_1}|$