

Numerical Analysis

gouziwu

April 14, 2019

Contents

1	Chap1 Mathematical Preliminaries	2
1.1	1.2 Roundoff Errors and Computer Arithmetic	2
1.2	1.3 ALgorithms and Convergence	2
2	Chap2 Solutions of equations in one variable	3
2.1	2.1 Bisection method	3
2.2	2.2 Fixed-Point Iteration	3
2.3	2.3 Newton's method	3
2.4	2.4 Error analysis for iterative methods	4
3	Chap3 Interpolation and polynomial approximation	6
3.1	3.1 Interpolation and the Lagrange polynomial	6
4	Chap6 Direct Methods for Solving Linear Systems	6
4.1	6.1 Linear Systems of Equations	6
4.2	6.2 Pivoting Strategies	6
4.3	6.5 Matrix Factorization	7
4.4	6.6 Special Types of Matrices	8
5	Chap7 Iterative techiniques in Matrix algebra	9
5.1	7.1 Norms of vectors and matrices	9
5.2	7.2 Eigenvalues and Eigenvectors	10
5.3	7.3 Iterative techniques for solving linear systems	10
5.4	7.4 Error bounds and iterative refinement	11
6	chap9 Approximating Eigenvalues	13
6.1	9.3 the power method	13

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 $y = 0.d_1d_2 \dots d_k d_{k+1}d_{k+2} \dots \times 10^n$, then

$$fl(y) = \begin{cases} 0.d_1d_2 \dots d_k \times 10^n & \text{chopping} \\ chop(y + 5 \times 10^{n-(k+1)}) = 0.\delta_1\delta_2 \dots \delta_k \times 10^n & \text{Rounding} \end{cases}$$

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 $\frac{|p - p^*|}{|p|} < 5 \times 10^{-t}$

$$\text{chopping } \left| \frac{y - fl(y)}{y} \right| = \left| \frac{0.d_1d_2 \dots d_k d_{k+1} \dots \times 10^n - 0.d_1d_2 \dots d_k \times 10^n}{0.d_1d_2 \dots d_k d_{k+1} \times 10^n} \right| = \left| \frac{0.d_{k+1} \dots}{0.d_1d_2 \dots} \right| \times 10^{-k} \leq \frac{1}{0.1} \times 10^{-k} = 10^{-k+1}$$

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

Finite digit arithmetic

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

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 E_n represents the magnitude of an error after n subsequent operations. If $E_n \approx C^n E_0$, 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 $\{\beta_n\}_{n=1}^\infty$, $\lim_{n \rightarrow \infty} \beta_n = 0$, $\{\alpha_n\}_{n=1}^\infty$, $\lim_{n \rightarrow \infty} \alpha_n = \alpha$. If a positive constant K exists with $|\alpha_n - \alpha| \leq K|\beta_n|$ for large n , then $\{\alpha_n\}_{n=1}^\infty$ converges to with **rate, or order, of convergence** $O(\beta_n)$.

Suppose $\lim_{h \rightarrow 0} G(h) = 0$, $\lim_{h \rightarrow 0} F(h) = L$ and $|F(h) - L| \leq K|G(h)|$ 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. *Bisection method* 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

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

2.2 2.2 Fixed-Point Iteration

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

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 $|g'(x)| \leq k$ 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 $p_n = g(p_{n-1})$, $n \geq 1$ converges to the unique point p in $[a, b]$.

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

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 $0 = f(p) \approx f(p_0) + f'(p_0)(p - p_0) \rightarrow p \approx p_0 - \frac{f(p_0)}{f'(p_0)}$ $p_n = p_{n-1} - \frac{f(p_{n-1})}{f'(p_{n-1})}$, for $n \geq 1$

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 $\{p_n\}$, $n \in \mathbb{N} \setminus \{0\}$ converging to p for any initial approximation $p \in [p - \delta, p + \delta]$.

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

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

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 $\alpha \geq 2$ s.t. $g \in C^\alpha[p - \delta, p + \delta]$, $g'(p) = \dots = g^{\alpha-1}(p) = 0$ and $g^\alpha(p) \neq 0$. Then the iterations with $p_n = g(p_{n-1})$, $n \geq 1$ is of *order α*

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

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

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

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

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

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

Proof.

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

□

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 $n \geq 1$ 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

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

□

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 $\lim_{x \rightarrow p} q(x) \neq 0$

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

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

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 $i = 0, \dots, n$ s.t. $L_{n,j}(x_j) = \delta_{ij}$. δ_{ij} Kronecker delta. Each $L_{n,i}$ has n roots $x_0, \dots, \hat{x}_i, \dots, x_n$. $L_{n,j}(x) = C_i(x - x_0) \dots (x - \hat{x}_i) \dots (x - x_n) = C_i \prod_{\substack{j=0 \\ j \neq i}}^n (x - x_j)$. $L_{n,j}(x_i) = 1 \rightarrow C_i = \prod_{j \neq i} \frac{1}{x_i - x_j}$. Hence $L_{n,i}(x) = \prod_{\substack{j=0 \\ j \neq i}}^n \frac{x - x_j}{x_i - x_j}$

Theorem 3.1. If x_0, x_1, \dots, 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 $a \leq x_0 < x_1 < \dots < x_n \leq b$ 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 $\Rightarrow R_n(x) = K(x) \prod_{i=1}^n (x - x_i)$. 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 $x_0 \dots x_n x$. Hence $g^{(n+1)}(\xi_x) = 0, \xi_x \in (a, b)$. $f^{(n+1)}(\xi_x) - P_n^{(n+1)}(\xi_x) - K(x)(n+1)! = R_n^{(n+1)}(\xi_x) - K(x)(n+1)!$. Thus $R_n(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!} \prod_{i=0}^n (x - x_i)$.

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

Partial Pivoting: Determine the smallest $p \geq k$ s.t. $|a_{pk}^{(k)}| = \max_{k \leq j \leq n} |a_{jk}^{(k)}|$ and interchange the p th and the k th rows

Scaled Partial Pivoting:

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

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

4.3 6.5 Matrix Factorization

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

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

Hence

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

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

$$A = LU$$

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 $i = 1, \dots, n$

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 it'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. $\sum |a_{ij}| \leq \max |a_{kk}|$; $(a_{ij})^2 < a_{ii}a_{jj}$ for each $i \neq j$
3. Each of A 's leading principal submatrices A_k has a positive determinant

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

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

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

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

5 Chap7 Iterative techniques in Matrix algebra

5.1 7.1 Norms of vectors and matrices

Definition 5.1. A *vector norm* on R^n is a function $\|\cdot\| : R^n \rightarrow \mathbb{R}$ with following properties for all $\mathbf{x}, \mathbf{y} \in R^n, \alpha \in C$

$$1. \|\mathbf{x}\| \geq 0; \|\mathbf{x}\| = 0 \iff \mathbf{x} = \mathbf{0}$$

$$2. \|\alpha\mathbf{x}\| = |\alpha| \cdot \|\mathbf{x}\|$$

$$3. \|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$$

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

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 $k \geq N(\epsilon)$

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 $\lim_{k \rightarrow \infty} \mathbf{x}_i^{(k)} = x_i$ for each $i = 1, 2, \dots, n$

Definition 5.3. If there exist positive constants C_1, C_2 s.t. $C_1 \|\mathbf{x}\|_B \leq \|\mathbf{x}\|_A \leq C_2 \|\mathbf{x}\|_B$. 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 $n \times n$:

$$1. \|\mathbf{A}\| \geq 0; \|\mathbf{A}\| = 0 \iff \mathbf{A} = \mathbf{0}$$

$$2. \|\alpha\mathbf{A}\| = |\alpha| \cdot \|\mathbf{A}\|$$

$$3. \|\mathbf{A} + \mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$$

$$4. \|\mathbf{AB}\| \leq \|\mathbf{A}\| \cdot \|\mathbf{B}\|$$

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

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

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

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 $\rho(A) \leq \|A\|$ for any natural norm

Proof. $|\lambda| \cdot \|\mathbf{x}\| = \|\lambda \mathbf{x}\| = \|\mathbf{Ax}\| \leq \|A\| \cdot \|\mathbf{x}\|$ □

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

5.3 7.3 Iterative techniques for solving linear systems

Jacobi iterative method.

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

In matrix form,

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

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

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

Gauss-Seidel iterative method

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

convergence of iterative methods

Theorem 5.4. *the following are equivalent:*

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

$$\begin{aligned}\mathbf{e}^{(k)} &= \mathbf{x}^{(k)} - \mathbf{x}^* = (T\mathbf{x}^{(k-1)} + \mathbf{c}) - (T\mathbf{x}^* + \mathbf{c}) = T(\mathbf{x}^{(k-1)} - \mathbf{x}^*) = \\ T\mathbf{e}^{(k-1)} &\Rightarrow \mathbf{e}^{(k)} = T^k \mathbf{e}^{(0)}. \quad \|\mathbf{e}^{(k)}\| \leq \|T\| \cdot \|\mathbf{e}^{(k-1)}\| \leq \dots \leq \|T\|^k \cdot \|\mathbf{e}^{(0)}\|\end{aligned}$$

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

5.4 7.4 Error bounds and iterative refinement

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

Theorem 5.6. *Suppose $\tilde{\mathbf{x}}$ is an approximation to the solution of $A\mathbf{x} = \mathbf{b}$ A is nonsingular matrix. Then for any natural norm,*

$$\|\mathbf{x} - \tilde{\mathbf{x}}\| \leq \|\mathbf{r}\| \cdot \|A^{-1}\|$$

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

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \|A\| \cdot \|A^{-1}\| \cdot \frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|}$$

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

Theorem 5.7. *If a matrix \mathbf{B} satisfies $\|\mathbf{B}\| < 1$ for some natural norm, then*

1. $\mathbf{I} \pm \mathbf{B}$ is nonsingular

2. $\|(\mathbf{I} \pm \mathbf{B})^{-1}\| \leq \frac{1}{1-\|\mathbf{B}\|}$

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

Hence $\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{\|\mathbf{A}^{-1}\| \cdot \|\delta\mathbf{A}\|}{1 - \|\mathbf{A}^{-1}\| \cdot \|\delta\mathbf{A}\|} = \frac{\|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\|}{1 - \|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\| \cdot \frac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|}}$

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

Theorem 5.8. *Suppose \mathbf{A} is nonsingular and $\|\delta\mathbf{A}\| \leq \frac{1}{\|\mathbf{A}^{-1}\|}$. The solution $\mathbf{x} + \delta\mathbf{x}$ to $(\mathbf{A} + \delta\mathbf{A})(\mathbf{x} + \delta\mathbf{x})$ approximates the solution \mathbf{x} of $\mathbf{A}\mathbf{x} = \mathbf{b}$ with the error estimate*

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

note:

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

iterative refinement:

Theorem 5.9. *Suppose \mathbf{x}^* is an approximation to the solution of $\mathbf{A}\mathbf{x} = \mathbf{b}$, \mathbf{A} is nonsingular matrix and $\mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x}$. Then for any natural norm, $\|\mathbf{x} - \mathbf{x}^*\| \leq \|\mathbf{r}\| \cdot \|\mathbf{A}^{-1}\|$, and if $\mathbf{x}, \mathbf{b} \neq \mathbf{0}$*

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

refinement

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

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 $|\lambda_1| > |\lambda_2| \geq \dots \geq |\lambda_n| \geq 0$

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

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

Note:

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

Rate of convergence. $\mathbf{x}^{(k)} = A\mathbf{x}^{(k-1)} = \lambda_1^k \sum_{j=1}^n \beta_j \left(\frac{\lambda_j}{\lambda_1}\right)^k \mathbf{v}_j$. Make $|\lambda_2/\lambda_1|$ as small as possible. Assume $\lambda_1 > \lambda_2 \geq \dots \geq \lambda_n, |\lambda_2| > |\lambda_n|$. 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 $|\lambda_1| \geq |\lambda_2| \geq \dots > |\lambda_n|$, then A^{-1} has $|\frac{1}{\lambda_n}| > |\frac{1}{\lambda_{n-1}}| \geq \dots \geq |\frac{1}{\lambda_1}|$