Numerical Linear Algebra

numerical integration

Theorem 0.1 (trapezoidal rule).

$$\int_a^b f(x)dx = \frac{h}{2}[f(x_0) + f(x_1)] - \frac{h^3}{12}f''(\xi)$$

Theorem 0.2 (Simpson's rule).

$$\int_{x_0}^{x_2} f(x)dx = \frac{h}{3}[f(x_0) + 4f(x_1) + f(x_2)] - \frac{h^5}{90}f^{(4)}(\xi)$$

the error term in Simpson's rule involves the fourth derivative of f, so it gives exact results when applied to any polynomial of degree three or less.

Definition 0.3 (degree of accuracy or precision). *the degree of accuracy of a quadrature formula is the largest positive integer n s.t. the formula is exact for x^k for each $k = 0, 1, 2, ..., n$*

Theorem 0.4 (newton-cotes formula).

composite numerical integration

Theorem 0.5 (composite Simpson's rule). *let $f \in C^4[a, b]$, n be even, $h = \frac{(b-a)}{n}$, and $x_j = a + jh$, for each $j = 0, 1, ..., n$. there exists a $\mu \in (a, b)$ for which the composite Simpson's rule for n subintervals can be written with its error term as*

$$\int_a^b f(x)dx = \frac{h}{3}\left[f(a) + 2\sum_{j=1}^{(n-2)-1} f(x_{2j}) + 4\sum_{j=1}^{n/2} f(x_{2j-1}) + f(b)\right]$$

$$-\frac{b-a}{180}h^4 f^{(4)}(\mu)$$

Theorem 0.6 (composite trapezoidal rule). *let $f \in C^2[a, b]$, $h = (b - a)/n$, and $x_j = a + jh$, for each $j = 0, 1, ..., n$, there exists a $\mu \in (a, b)$ for which the composite trapezoidal rule for n subintervals can be written with its error term as*

$$\int_a^b f(x)dx = \frac{h}{2}\left[f(a) + 2\sum_{j=1}^{n-1} f(x_j) + f(b)\right] - \frac{b-a}{12}h^2 f''(\mu)$$

Theorem 0.7 (composite midpoint rule). *let $f \in C^2[a, b]$, b be even, $h = (b - a)/(n + 2)$, and $x_j = a + (j + 1)h$ for each $j = -1, 0, ..., n + 1$. there exists a $\mu \in (a, b)$ for which the composite midpoint rule for $n + 2$ subintervals can be written with its error term as*

$$\int_a^b f(x)dx = 2h\sum_{j=0}^{n/2} f(x_{2j}) + \frac{b-a}{6}h^2 f''(\mu)$$

Theorem 0.8 (Romberg integration). *the composite trapezoidal rule can be written with an error form:*

$$\int_a^b f(x)dx = \frac{h}{2}\left[f(a) + 2\sum_{j=1}^{n-1}(x_j) + f(b)\right] + K_1h^2 + K_2h^4 + ...$$

where each K_I is a constant that depends only on $f^{(2i-1)}(a)$ and $f^{(2i-1)}(b)$. we combine this trapezoidal rule and the Richardson extrapolation and get the Romberg integration:

1. *first we use composite trapezoidal rule to calculate the case when $n = 1, 2, 4, 8, 16, ...$, denoted by $R_{1,1}, R_{2,1}, R_{3,1}, R_{4,1}, ...$*
2. *we calculate* $R_{k,1} = \frac{1}{2}\left[R_{k-1,1} + h_{k-1}\sum_{i=1}^{2^{k-2}} f(a + (2i - 1)h_k)\right]$
3. *we calculate $R_{k,j}$ using the formula:*

   $$R_{k,j} = R_{k,j-1} + \frac{1}{4^{j-1} - 1}(R_{k,j-1} - R_{k-1,j-1}),$$

   $$\text{for } k = j, j + 1, ...$$

4. *recursively calculate $R_{k,j}$ to generalize a table:*

k	$O(h_k^2)$	$O(h_k^4)$	$O(h_k^6)$	$O(h_k^{2n})$
1	$R_{1,1}$			
2	$R_{2,1}$	$R_{2,2}$		
$\vdots$	$\vdots$	$\vdots$	$\ddots$	
n	$R_{n,1}$	$R_{n,2}$	$\cdots$	$R_{n,n}$

Gaussian quadrature

Definition 0.9 (legendre polynomials). *Legendre polynomials are a collection $\{P_0(x), P_1(x), ..., P_n(x)\}$ with properties:*

1. *for each n, $P_n(x)$ is a monic polynomial of degree n*
2. $\int_{-1}^{1} P(x)P_n(x)dx = 0$ *whenever $P(x)$ is a polynomial of degree less than n*

the roots of these polynomials are distinct, lie in the interval $(-1, 1)$, have a symmetry with respect to the origin, and, most importantly, are the correct choice for determining the parameters that give us the nodes and coefficients for our quadrature method.

Theorem 0.10. *suppose that $x_1, x_2, ..., x_n$ are the roots o the n-th Legendre polynomial $P_n(x)$ and that for each $i = 1, 2, ..., n$, the numbers c_i are defined by*

$$c_i = \int_{-1}^{1}\prod_{j=1,j\neq i}^{n}\frac{x - x_j}{x_i - x_j}dx$$

if $P(x)$ is any polynomial of degree less than $2n$, then

$$\int_{-1}^{1} P(x)dx = \sum_{i=1}^{n} c_iP(x_i)$$

Proof. □

Theorem 0.11 (gaussian quadrature on arbitrary intervals). *an integral $\int_a^b f(x)dx$ over an arbitrary $[a, b]$ can be transformed into an integral over $[-1, 1]$ by using the change of variables:*

$$t = \frac{2x - a - b}{b - a} \leftrightarrow x = \frac{1}{2}[(b - a)t + a + b]$$

this permits gaussian quadrature to be applied to any interval $[a, b]$ because

$$\int_a^b f(x)dx = \int_{-1}^{1} f\left(\frac{(b-a)t + (b+a)}{2}\right)\frac{(b-a)}{2}dt$$

Direct Methods for Solving Linear Systems

Matrix Factorization

Theorem 0.12 (Gaussian elimination). *gaussian elimination algorithm:*

1. *for $k = 1, 2, ..., n - 1$:*
2. *for $i = k + 1, k + 2, ..., n$*
3. $m_{ik} = \frac{a_{ik}}{a_{kk}}$
4. *for $j = k + 1, ..., n + 1$*
5. $a_{ij} = a_{ij} - m_{ik}a_{kj}$
   to find x_i we use backward substitution:
6. *for $i = n, n - 1, ..., 1$*
7. $x_i = \frac{a_{i,n+1} - \sum_{j=n+1}^{n} a_{ij}x_{ij}}{a_{ii}}$

we use pivoting strategies to avoide dividing by zero:

Pivoting Strategies

Definition 0.13 (partial pivoting)**.** *difficulties can arise when the pivot element $a_{kk}^{(k)}$ is small relative to the entries $a_{ij}^{(k)}$, for $k \leq i \leq n$ and $k \leq j \leq n$. To avoid this problem, pivoting is performed by selecting an element $a_{pq}^{(k)}$ with a larger magnitude as the pivot and interchanging the k-th and p-th rows. This can be followed by the interchange of the k-th and q-th columns, if necessary.*
The simplest strategy, called partial pivoting, is to select an element in the same column that is below the diagonal and has the largest absolute value; specifically, we determine the smallest $p \geq k$ such that

$$|a_{pk}^{(k)}| = \max_{k \leq i \leq n} |a_{ik}^{(k)}|$$

and perform $(E_k) \leftrightarrow (E_p)$. in this case, no interchange of columns is used.

Definition 0.14 (scaled partial pivoting)**.** *it places the element in the pivot position that is largest relative to the entries in its row. The first step in this procedure is to define a scale factor s_i for each row as*

$$s_i = \max_{1 \leq j \leq n} |a_{ij}|$$

if we have $s_i = 0$ for some i, then the system has no unique solution since all entries in the i-th row are 0. assuming that this is not the case, the appropriate row interchange to place zeros in the first column is determined by choosing the least integer p with

$$\frac{|a_{p1}|}{s_p} = \max_{1 \leq k \leq n} \frac{|a_{k1}|}{s_k}$$

and performing $(E_1) \leftrightarrow (E_p)$. the effect of scaling is to ensure that the largest element in each row has a relative magnitude of 1 before the comparison for row interchange is performed. in a similar manner, before eliminating the variable x_i using the operations

$$E_k = m_{ki}E_i, \text{ for } k = i+1, ..., n$$

we select the smallest integer $p \geq i$ with

$$\frac{|a_{pi}|}{s_p} = \max_{i \leq k \leq n} \frac{|a_{ki}|}{s_k}$$

Definition 0.15 (permutation matrix)**.** *a permutation matrix combines a branch of orthonormal vectors in a certain unique and P^TP is the matrix dot product of these orthonormal vectors.*

we summarize the Gaussian elimination with partial pivoting:

1. we permute the rows of A: $P^TA = LU$, P is a permutation matrix, which combines a branch of orthonormal vectors in a certain sequence and P^TP is the matrix dot product of these orthonormal vectors. We have
$$A = PP^TA = PLU$$
2. if we do total pivoting: we factor
$$A = PLUQ$$
where P and Q are suitable permutation matrices

Definition 0.16 (the number of flops)**.** *which usually means the number of multiplications and divisions. we do not count additions and subtractions or other operations because the time of implementing these operations (multiplication and division) is proportional to the number of flops.*

Definition 0.17 (Cholesky-factorization)**.** *$A = LL^T$, L is a lower triangular matrix.*

Special Types of Matrices

Definition 0.18 (diagonally dominant matrix)**.** *the $n \times n$ matrix is said to be diagonally dominant when*

$$|a_{ii}| \geq \sum_{j=1, j\neq i}^{n} |a_{ij}|$$

a diagonally dominant matrix is said to be strictly diagonally dominant when the inequality in the above equality is strict for each n, that is, when

$$|a_{ii}| > \sum_{j=1, j\neq i}^{n} |a_{ij}|$$

Definition 0.19 (positive definite matrix)**.** *a matrix A is positive definite if it is symmetric and $x^TAx > 0$ for all vectors $x \neq 0$*

Proposition 0.20. *some facts about positive definite matrices:*

1. *the diagonal entries of a p.d. matrix are positive*
2. *the eigenvalues of a p.d. matrix are positive*
3. *a symmetric matrix whose eigenvalues are all positive is p.d.*
4. *the determinant of a p.d. matrix is positive*
5. *a symmetric matrix may have a positive determinant but not be p.d.*

Theorem 0.21. *if A is positive definite, the Cholesky decomposition always exists*

Proof. we give two proofs for the theorem □

Definition 0.22 (band matrix; tridiagonal matrix)**.** *a matrix $A = (a_{ij})_{n\times n}$ is called a band matrix if there are two integers $1 < p, q < n$, s.t.*

$$a_{ij} = 0, \quad \text{for } j - i \geq p \text{ or } j - i \leq -q$$

or equivalently,

$$a_{ij} \begin{cases} \neq 0 & \text{if } -q < j - i < p \\ = 0 & \text{otherwise} \end{cases}$$

the bandwidth of the above matrix is defined as $w = p + q - 1$. if $p = q = 2$, the matrix is called tridiagonal.

Proposition 0.23 (LU-decomposition for tridiagonal matrix; Crout decomposition)**.** *for a tridiagonal matrix A, it can be factored into the product of lower triangular L and unit upper triangular U, where*

$$L = \begin{bmatrix} l_{11} & 0 & 0 & \cdots & \cdots & 0 \\ l_{21} & l_{22} & 0 & \ddots & \ddots & \vdots \\ 0 & l_{32} & l_{33} & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & l_{n,n-1} & l_{n,n} \end{bmatrix}$$

$$U = \begin{bmatrix} 1 & u_{12} & 0 & \cdots & \cdots & 0 \\ 0 & 1 & u_{23} & \ddots & \ddots & \vdots \\ 0 & 0 & 1 & u_{34} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \ddots & \ddots & u_{n-1,n} \\ 0 & \cdots & \cdots & 0 & 0 & 1 \end{bmatrix}$$

the number of operations of the Crout factorization is $O(n)$ for solving a tridiagonal linear system.

Iterative Techniques in Matrix Algebra

Norms of Vectors and Matrices

Definition 0.24 (vector norm)**.** *a function $f : \mathbb{R}^n \to \mathbb{R}^n$ that associates a number $\|x\|$ with a vector x is called a vector norm if the following four properties hold for all vectors x, y and scalars k:*

1. $\|x\| \geq 0$
2. $\|x\| = 0 \to x = 0$
3. $\|kx\| = |k|\|x\|$
4. $\|x + y\| \leq \|x\| + \|y\|$

Definition 0.25 (1-norm; 2-norm; $\infty-$norm; weighted p-norm)**.** *for the vector $x = (x_1, x_2, ..., x_n)^t$, we have:*

1. $\|X\|_1 = \sum_{i=1}^n |x_i|$
2. $\|X\|_2 = (\sum_{i=1}^n x_i^2)^{1/2}$
3. $\|x\|_\infty = \max_{1\leq i \leq n} |x_i|$
4. $\omega \in \mathbb{R}^n, \omega_i > 0, i = 0, 1, ..., n$
$\|x\|_{\omega,p} = (\sum_{i=1}^n \omega_i |x_i|^p)^{1/p}$

Theorem 0.26 (Hölder inequality). $|x^T y| \leq \|x\|_p \|y\|_q$ *given* $\frac{1}{p} + \frac{1}{q} = 1$

Definition 0.27 (matrix norm). *the space of $m \times n$ matrices is isomorphic to $\mathbb{R}^{m \times n}$ and we can apply the same definition as we did for vectors: the function $A : \mathbb{R}^{m \times n} \to \mathbb{R}$ that associates a number $\|A\|$ with a matrix A is called a matrix norm if we have:*

1. $\|A\| \geq 0$
2. $\|A\| = 0$, *if and only if A is O, the matrix with all 0 entries*
3. $\|\alpha A\| = |\alpha| \|A\|$
4. $\|A + B\| \leq \|A\| + \|B\|$

hold for all $m \times n$ matrices A, B and scalars k

Theorem 0.28 (product property). $\|AB\| \leq \|A\| \|B\|$ *since* $||Ax|| = \frac{||Ax||}{||x||} ||x|| \leq ||A|| \cdot ||x||$

Definition 0.29 (Frobenius norm).
$\|A\|_F = \sqrt{\sum_i \sum_j a_{ij}^2} = \sqrt{tr(A^t A)}$ *using the fact: let $B = A^t A$, $tr(B) = \sum_i b_{ii}$ and $b_{ii} = \sum_j a_{ji} a_{ji} = \sum_j a_{ji}^2$*

Definition 0.30 (induced matrix norm). *given vector norms in $\mathbb{R}^m$ and $\mathbb{R}^n$, we define the induced matrix norm of an $m \times n$ matrix A as*

$$||A|| = \max_{\|x\|=1} \|Ax\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|}$$

Exercise 0.31. *about matrix norm, we have the following problems:*

1. *the matrix norm $\| \cdot \|_1$ for matrix $A = (a_{ij})_{m \times n} \in \mathbb{R}^{m \times n}$ is defined by $\|A\|_1 = \max_{\|x\|_1 = 1} ||Ax||_1$, we can show that $\|A\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^m |a_{ij}|$, which is the max among all absolute value sums of each row.*
2. *the matrix norm $\| \cdot \|_2$ for matrix $A = (a_{ij})_{n \times n} \in \mathbb{R}^{n \times n}$ is defined by $\|A\|_2 = \max_{\|x\|_2 = 1} \|Ax\|_2$, we can show that*
$$\|A\|_2 = \sqrt{\rho(A^T A)}$$
3. *the matrix norm $\| \cdot \|_\infty$ for matrix $A = (a_{ij})_{m \times n} \in \mathbb{R}^{m \times n}$ is defined by $\|A\|_\infty = \max_{\|x\|_\infty = 1} ||Ax||_\infty$, we can show that $\|A\|_\infty = \max_{1 \leq i \leq m} \sum_{j=1}^n |a_{ij}|$, which is the max among all absolute value sums of each column.*

The Jacobi and Gauss-Siedel Iterative Techniques

we are going to solve the problem: $Ax = b$, we define $A = D - L - U$.

Definition 0.32 (Jacobi's method).

$$Dx = (L + U)x + b \to X = D^{-1}(L + U)x + D^{-1}b$$

let $T_j = D^{-1}(L + U)$ and $c_j = D^{-1}b$, we have

$$x^{(k)} = T_j x^{(k-1)} + c_j$$

Definition 0.33 (Gauss-Siedel's method).

$$(D - L)x^{(k)} = Ux^{(k-1)} + b$$

let $T_g = (D - L)^{-1}U$ and $c_g = (D - L)^{-1}b$, we have

$$x^{(k)} = T_g x^{(k-1)} + c_g$$

general iteration methods

we analyze the formula

$$x^{(k)} = Tx^{k-1} + c, k = 1, 2, ...$$

where $x^{(0)}$ is arbitrary. we want to prove the theorem: for any $x^{(0)} \in \mathbb{R}^n$, the sequence $\{x^{(k)}\}_{k=0}^\infty$ defined by

$$x^{(k)} = Tx^{(k-1)} + c, \quad k = 1, 2, ...$$

converges to the unique solution of $x = Tx + c$ if and only if $\rho(T) < 1$

Definition 0.34 (spectral radius). *the spectral radius $\rho(A)$ of a matrix A is defined by*

$$\rho(A) = \max |\lambda|$$

where the maximum is taken among all the eigenvalues of A.

Theorem 0.35. *if A is an $n \times n$ matrix, then*

1. $\|A\|_2 = [\rho(A^T A)]^{1/2}$
2. $\rho(A) \leq \|A\|$ *for any natural norm* $\| \cdot \|$

Theorem 0.36. *for any square matrix A and any $\epsilon > 0$, there exists a natural norm $\| \cdot \|$ s.t.*

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

Definition 0.37 (convergent matrix). *We call an $n \times n$ matrix A convergent if*

$$lim_{k \to \infty} (A^k)_{ij} = 0, for\ each\ i = 1, 2, ..., n\ and\ j = 1, 2, ..., n.$$

Theorem 0.38. *the following statements are equivalent:*

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

Lemma 0.39. *if the spectral radius satisfies $\rho(T) < 1$, then $(I - T)^{-1}$ exists, and*

$$(I - T)^{-1} = I + T + T^2 + ... = \sum_{k=0}^\infty T^k$$

Theorem 0.40. *for any $x^{(0)} \in \mathbb{R}^n$, the sequence $\{x^{(k)}\}_{k=0}^\infty$ defined by*

$$x^{(k)} = Tx^{(k-1)} + c, \quad k = 1, 2, ...$$

converges to the unique solution of $x = Tx + c$ if and only if $\rho(T) < 1$

Theorem 0.41. *if A is strictly diagonally dominant, then for any choice of $x^{(0)}$, both the Jacobi and Gauss-Seidel methods give sequence $\{x^{(k)}\}_{k=0}^\infty$ that converge to the unique solution of $Ax = b$*

Relaxation Techniques for Solving Linear Systems

Definition 0.42 (residual vector). *suppose $\tilde{x} \in \mathbb{R}^n$ is an approximation to the solution of the linear system defined by $Ax = b$, the residual vector for $\tilde{x}$ with respect to the system is $r = b - A\tilde{x}$*

Remark 0.43. *the relation between the residual vectors and the G-S method:*
using G-S method, the m-th component of $r_i^{(k)}$ is

$$r_{mi}^{(k)} = b_m - \sum_{j=1}^{i-1} a_{mj} x_j^{(k)} - \sum_{j=i+1}^{n} a_{mj} x_j^{(k-1)} - a_{mi} x_i^{(k-1)}$$

so we have

$$a_{ii} x_i^{(k-1)} + r_{ii}^{(k)} = b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k-1)}$$

in G-S method,

$$x_i^{(k)} = \frac{1}{a_{ii}} \left[b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k-1)} \right]$$

so we have

$$a_{ii} x_i^{(k-1)} + r_{ii}^{(k)} = a_{ii} x_i^{(k)}$$

G-S method can be characterized as choosing $x_i^{(k)} = x_i^{(k-1)} + \frac{r_{ii}^{(k)}}{a_{ii}}$ G-S method is characterized by choosing each $x_{i+1}^{(k)}$ in a way that the i-th component $r_{i+1}^{(k)}$ is zero. If we modify the G-S procedure by a coefficient ω:

$$x_i^{(k)} = x_i^{(k-1)} + \omega \frac{r_{ii}^{(k)}}{a_{ii}}$$

we usually let $\omega > 1$, and the improved method is called SOR method (Successive over-relaxation)

the formula in vector form of SOR method is

$$(D - \omega L)x^{(k)} = [(1 - \omega)D + \omega U]x^{(k-1)} + \omega b$$

we have

$$T_\omega = (D - \omega L)^{-1}[(1 - \omega)D + \omega U] \quad c_\omega = \omega(D - \omega L)^{-1}b$$

Remark 0.44. the idea of SOR method is $x_i^{(k+1)} = (1-\omega)x_i^{(k)} + \omega\tilde{x}_i^{(k+1)}$, where $\tilde{x}_i^{(k+1)}$ is by G-S method:

$$\tilde{x}_i^{(k+1)} = \left(b_i - \sum_{j=1}^{i-1} a_{ij}{x_j}^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)} \right) / a_{ii}$$

Theorem 0.45 (Kahan)**.** if $a_{ii} \neq 0$, for each $i = 1, 2, ..., n$, then $\rho(T_\omega) \geq |\omega - 1|$, this implies that the SOR method can converge only if $0 < \omega < 2$

Theorem 0.46 (Ostrowski-Reich)**.** if A is a positive definite matrix and $0 < \omega < 2$, then the SOR method converges for any choice of initial approximate vector $x^{(0)}$

Proof. □

Theorem 0.47. 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 $\rho(T_\omega) = \omega - 1$

Error Bounds and Iterative Refinement

Definition 0.48 (condition number)**.** the condition number of the nonsingular matrix A relative to a norm $\|\cdot\|$ is

$$K(A) = \|A\| \cdot \|A^{-1}\|$$

A matrix A is well conditioned if $K(A)$ is close to 1, and is ill conditioned when $K(A)$ is significantly greater than 1.

Remark 0.49. let λ_{max} and λ_{min} denote the largest and the smallest of the absolute values of the eigenvalues of A, since we have: for any induced matrix norm, $\|A\| \geq \lambda_{max}$ and $\|A^{-1}\| \geq \frac{1}{\lambda_{min}}$, we have

$$\|A\| \cdot \|A^{-1}\| \geq \frac{\lambda_{max}}{\lambda_{min}} \geq 1$$

Remark 0.50. we know that a matrix is singular if and only if zero is one of its eigenvalues. consider a non-singular matrix A where one eigenvalue remains constant, and another approaches zero. In that sense, A approaches singularity and the condition number approaches infinity.

in the real applications, the system $Ax = b$ are usually not exact, but with small perturbations, so the system becomes

$$(A + \delta A)x = b + \delta b$$

we hope that if $\|\delta A\|$ and $\|\delta b\|$ are small, they should yield a solution $\tilde{x}$ for which $\|\tilde{x} - x\|$ is correspondingly small. For some ill-conditioned system, the hypothesis is not true. We have the following theorems for the error estimation.

Theorem 0.51. suppose that $\tilde{x}$ is an approximation to the solution of $Ax = b$, A is a nonsingular matrix, and r is the residual vector for $\tilde{x}$. then, for any natural norm, $\|\tilde{x} - x\| \leq \|A^{-1}\| \cdot \|r\|$ and if $x \neq 0$ and $b \neq 0$,

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

in the real applications, the matrix A and the right hand side are usually not exact, but with small perturbations. so the system $Ax = b$ becomes $(A + \delta A)x = b + \delta b$

Theorem 0.52. suppose that A is nonsingular and

$$||\delta A|| \leq \frac{1}{||A^{-1}||}$$

The solution $\tilde{x}$ to $(A + \delta A)x = b + \delta b$ approximates the solution x of $Ax = b$ with the error estimate

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

Proof. □

Lemma 0.53. if $\|B\| < 1$, then $I \pm B$ is nonsingular, and

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

Proof. □

QR factorization

Definition 0.54 (orthogonal matrix)**.** a matrix Q is called orthogonal if its columns $\{q_1^t, q_2^t, ..., q_n^t\}$ form an orthogonal set in $\mathbb{R}^n$

Theorem 0.55. suppose that Q is an orthogonal $n \times n$ matrix, then

1. Q is invertible with $Q^{-1} = Q^t$
2. for any x and y in $\mathbb{R}^n$, $(Qx)^t Qy = x^t y$
3. for any x in $\mathbb{R}^n$, $\|Qx\|_2 = \|x\|_2$
4. any invertible matrix Q with $Q^{-1} = Q^t$ is orthogonal

Theorem 0.56. a symmetric matrix A is positive definite if and only if all the eigenvalues of A are positive.

Proof. suppose that A is positive definite and that λ is an eigenvalue of A with associated eigenvector x, with $\|x\|_2 = 1$. Then

$$0 < x^t Ax = \lambda x^t x = \lambda\|x\|_2 = \lambda$$

□

Householder's method

Definition 0.57 (householder transformation)**.** let $u \in \mathbb{R}^n$ with $u^t u = 1$. the $n \times n$ matrix

$$P = I - 2uu^t$$

is called a Householder transformation

Theorem 0.58. a Householder transformation, $P = I - 2uu^t$, is symmetric and orthogonal, so $P^{-1} = P$

we use Householder's method to zero a column a of A: we want to find u s.t. $Ha = \alpha e$. write $u = \frac{v}{\|v\|_2}$, we have

$$Ha = (I - 2uu^T)a = a - (2u^T a)u = a - (2u^T a)\frac{v}{\|v\|_2}$$

so we have

$$\alpha e = a - \gamma v$$

we rewrite the equation as

$$v = \beta a + \alpha e$$

since α and β are parameters, then we have $u = \frac{v}{\|v\|_2}$ and plug it into the equation $Ha = (I - 2uu^T)a = a - (\frac{2(\beta a + \alpha e)^T a}{\beta^2 a^T a + 2\beta\alpha a_1 + \alpha^2})(\beta\alpha + \alpha e) = g \cdot a + h \cdot e$, we know that $Ha = \alpha e$, so $g \cdot a = 0 \Rightarrow g = (1 - \frac{2(\beta a + \alpha e)^T a}{\beta^2 a^T a + 2\alpha\beta a_1 + \alpha^2}) = 0$, where a_1 is the 1st entry of a, solve the equation, we find $\beta = 1$ and $\alpha = \pm\|a\|_2$ is a solution, so we let $u = \frac{a \pm \|a\|_2 e}{\|a \pm \|a\|_2 e\|}$, to make the first entry of a as large as possible, the usual choice is $u = \frac{a + sign(a_1)\|a\|_2 e}{\|a + sign(a_1)\|a\|_2 e\|}$

Remark 0.59. recall the property that $Ha = \alpha e$, so A multiplying H makes a matrix HA whose entries in the first column are all zero except the first entry HA_{11}. we repeat the above process towards HA to get H' which gives $H'HA$, and finally we get $H_n H_{n-1} ... H_1 A = R$, R is upper triangular, since H_i are all orthogonal, we let $Q = (H_n H_{n-1} ... H_1)^T$ is orthogonal and have $A = Q(Q^T A) = Q(H_n H_{n-1} ... H_1)A = QR$

eigenvalue problem

power method

Definition 0.60 (dominant eigenvalue)**.**

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

then we call λ_1 the dominant eigenvalue.

power method is used to find λ_1 and its corresponding eigenvector x.

Theorem 0.61. to find the dominant eigenvalue of a square matrix A, we first randomly pick a vector $q^{(0)}$, for $k = 0, 1, ...,$ let $z^{(k+1)} = Aq^{(k)}$, $q^{(k)} = \frac{z^{(k+1)}}{\|z^{(k+1)}\|_\infty}$, then we have $x = \lim_{k\to\infty} z^{(k)}$

Proof. □

Remark 0.62. *to get the corresponding eigenvalue λ_1, we note that λ_1 is the value that minimize $F(\lambda) = \|Aq - \lambda q\|_2^2$, so we have $F'(\lambda) = 0 \to \lambda_1 = \frac{x^T(A^T+A)x}{2x^Tx}$. if A is symmetric, we have $\lambda_1 = \frac{q^TAq}{q^Tq}$, which is called Rayleigh Quotient*

review

1. prove trapezoidal rule, simpson‘s rule
2. prove LL-decomposition
3. prove $||x||_\infty = \max_{||x||=1} ||Ax||_\infty = \max_{i=1,2,\ldots,m} \sum_{j=1}^{n} |a_{ij}|$: the max among all absolute value sums of each column
4. prove $||A||_2 = [\rho(A^tA)]^{1/2}, \rho(A) \leq \|A\|$ for any natural norm
5. prove $\rho(A) \leq ||A||$ for any natural norm $|| \cdot ||$
6. prove 5 equivalent propositions of convergent matrix
7. prove $\rho(T) < 1$, then $(I - T)^{-1}$ exists
8. prove $\rho(T) < 1 \leftrightarrow$ the sequence $\{x^{(k)}\}$ defined by $x^{(k)} = Tx^{(k-1)} + c$ converges to the unique solution of $x = Tx + c$
9. prove ostrowski-reich theorem
10. prove hilbert matrix is positive definite
11. prove error estimate theorem and lemma in week13