A FIRST COURSE

IN

NUMERICAL ANALYSIS

A FIRST COURSE

IN

NUMERICAL ANALYSIS

MAT4001 Notebook

Prof. Yutian Li

The Chinese University of Hongkong, Shenzhen

香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen

Contents

Acknowledgments

This book is from the MAT4001 in fall semester, 2018.

CUHK(SZ)

Notations and Conventions

$\mathbb{R}^n$	n-dimensional real space
$\mathbb{C}^n$	n-dimensional complex space
$\mathbb{R}^{m\times n}$	set of all $m \times n$ real-valued matrices
$\mathbb{C}^{m\times n}$	set of all $m \times n$ complex-valued matrices
x_i	ith entry of column vector $\boldsymbol{x}$
a_{ij}	(i,j)th entry of matrix $\boldsymbol{A}$
$\boldsymbol{a}_i$	ith column of matrix $\boldsymbol{A}$
$\boldsymbol{a}_i^{\mathrm{T}}$	ith row of matrix $\boldsymbol{A}$
$\mathbb{S}^n$	set of all $n \times n$ real symmetric matrices, i.e., $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ and $a_{ij} = a_{ji}$ for all i,j
$\mathbb{H}^n$	set of all $n \times n$ complex Hermitian matrices, i.e., $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ and $\bar{a}_{ij} = a_{ji}$ for all i,j
$\boldsymbol{A}^{\mathrm{T}}$	transpose of $\boldsymbol{A}$, i.e, $\boldsymbol{B} = \boldsymbol{A}^{\mathrm{T}}$ means $b_{ji} = a_{ij}$ for all i,j
$\boldsymbol{A}^{\mathrm{H}}$	Hermitian transpose of $\boldsymbol{A}$, i.e, $\boldsymbol{B} = \boldsymbol{A}^{\mathrm{H}}$ means $b_{ji} = \bar{a}_{ij}$ for all i,j
$\mathrm{trace}(\boldsymbol{A})$	sum of diagonal entries of square matrix $\boldsymbol{A}$
$\mathbf{1}$	A vector with all 1 entries
$\mathbf{0}$	either a vector of all zeros, or a matrix of all zeros
$\boldsymbol{e}_i$	a unit vector with the nonzero element at the ith entry
$\mathcal{C}(\boldsymbol{A})$	the column space of $\boldsymbol{A}$
$\mathcal{R}(\boldsymbol{A})$	the row space of $\boldsymbol{A}$
$\mathcal{N}(\boldsymbol{A})$	the null space of $\boldsymbol{A}$
$\mathrm{Proj}_{\mathcal{M}}(\boldsymbol{A})$	the projection of $\boldsymbol{A}$ onto the set $\mathcal{M}$

Chapter 1

Week1

1.1. Differentiation

Definition 1.1 [Forward and Backforward Difference Formula]

$$f'(x_0) = \frac{f(x_0+h) - f(x_0)}{h} - \frac{h}{2} f''(\xi)$$

Consider the Taylor expansion, we have

$$f''(x_0) = \frac{1}{h^2}[f(x_0 - h) - 2f(x_0) + f(x_0+h)] - \frac{h^2}{12} f^{(4)}(\xi)$$

Definition 1.2 [General Derivative Approximation] The interpolation formula gives

$$f(x) = \sum_{k=0}^{n} f(x_k) L_k(x) + \frac{(x-x_0)\cdots(x-x_n)}{(n+1)!} f^{(n+1)}(\xi_x)$$

Differentiating both sides gives (x_j is one of the node points)

$$f'(x_j) = \sum_{k=0}^{n} f(x_k) L_k'(x_j) + \frac{1}{(n+1)!} f^{(n+1)}(\xi_{x_j}) \prod_{k=0, k\neq j}^{n} (x_j - x_k)$$

R Three-point formula:

$$f'(x_j) = f(x_0)\left[\frac{2x_j - x_1 - x_2}{(x_0 - x_1)(x_0 - x_2)}\right] + f(x_1)\left[\frac{2x_j - x_0 - x_2}{(x_1 - x_0)(x_1 - x_2)}\right] + f(x_2)\left[\frac{2x_j - x_0 - x_1}{(x_2 - x_0)(x_2 - x_1)}\right] + \frac{1}{6}f^{(3)}(\xi_j)\prod_{k=0,k\neq j}^{2}(x_j - x_k)$$

Substituting $x_j = x_0, x_1, x_2$, we obtain:

$$f'(x_0) = \frac{1}{2h}[-3f(x_0) + 4f(x_0 + h) - f(x_0 + 2h)] + \frac{h^2}{3}f^{(3)}(\xi_0)$$
$$f'(x_0) = \frac{1}{2h}[-f(x_0 - h) + f(x_0 + h)] - \frac{h^2}{6}f^{(3)}(\xi_1)$$

Definition 1.3 [Richardson Extrapolation] For the approximation with the form

$$M = N_1(h) + K_1h^2 + K_2h^4 + \cdots$$

we have the approximation

$$N_j(h) = N_{j-1}(\frac{h}{2}) + \frac{N_{j-1}(h/2) - N_{j-1}(h)}{4^{j-1} - 1}$$

where $N_j(h)$ has order $O(h^{2j})$

round-off error in $N_1(h/2^k)$; we recommend comparing the final diagonal entries to ensure accuracy.

Definition 1.4 [Quadrature formula]

$$\int_a^b f(x)\,\mathrm{d}x \approx \sum_{i=0}^{n} a_i f(x_i)$$

where

$$
\begin{aligned}
a_i &= \int_a^b L_i(x)\,\mathrm{d}x \\
E(f) &= \frac{1}{(n+1)!}\int_a^b \prod_{i=0}^{n}(x-x_i) f^{(n+1)}(\xi_x)\,\mathrm{d}x
\end{aligned}
$$

1. Trapezoidal Rule:

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

Seperating into subintervals $[x_{k-1}, x_k]$, we have, with $h = (b-a)/n$,

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

2. Simpson's rule: for $h = (x_2 - x_0)/2$,

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

Error Bound: Taylor expansion of $f(x)$, and bound the integrand

$$\frac{1}{24}\int_{x_0}^{x_2} f^{(4)}(\xi_x)(x-x_1)^4 = \frac{f^{(4)}(\xi_1)}{120}(x-x_1)^5\bigg|_{x_0}^{x_2} = \frac{f^{(4)}(\xi_1)}{60}h^5$$

Seperating into subintervals $[x_0, x_2], [x_2, x_4], \ldots, [x_{n-2}, x_n]$, we have, with $h = (b - a)/n$,

$$
\begin{aligned}
\int_a^b f(x)\,\mathrm{d}x &= \sum_{j=1}^{n/2}\left\{\frac{h}{3}[f(x_{2j-2} + 4f(x_{2j-1}) + f(x_{2j}))]\right\} \\
\text{Error} &= -\frac{h^5}{90}\sum_{j=1}^{n/2} f^{(4)}(\xi_j) = -\frac{h^5}{90}\frac{n}{2}f^{(4)}\mu = -\frac{b-a}{180}h^4 f^{(4)}(\mu)
\end{aligned}
$$

3. Composite Mid-point rule: for subintervals $[x_{-1}, x_1], \ldots, [x_{n-1}, x_{n+1}]$ with centers

$x_0, x_2, \dots, x_n$, and $h = (b-a)/(n+2)$

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

Definition 1.5 [Degree of Precision] The degree of precision of a quadrature formula is n if and only if the error is zero for all polynomials of degree $k = 0, 1, \cdots, n$, but is not zero for some polynomial of degree $n+1$.

Definition 1.6 [Romberg Method] First column: for $k = 2, \dots, n$, $h_k = (b-a)/(2^{k-1})$

$$R_{k,1} = \frac{1}{2}\left[R_{k-1,1} + h_{k-1}\sum_{i=1}^{2^{k-2}} f(a + (2i-1)h_k)\right]$$

For $k = j, j+1, \dots$,

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

Stopping Criteria: $|R_{n-1,n-1} - R_{n,n}| < \text{tol}$ and $|R_{n-2,n-2} - R_{n-1,n-1}| < \text{tol}$. Ensure two differently generated sets of approximations agree within the specified tolerance.

Definition 1.7 [Gauss-Quadrature Rule] Generate Legendre poylnomial $P_0(x) = 1, P_1(x) = x$,

$$(n+1)P_{n+1}(x) = (2n+1)xP_n(x) - nP_{n-1}(x)$$

Let $\{x_0, \dots, x_n\}$ be rooots of $P_{n+1}(x)$, and

$$w_i = \int_{-1}^{1} l_i(x)\,\mathrm{d}x = \int_{-1}^{1}\prod_{j=0, j\neq i}^{n}\frac{x-x_j}{x_i-x_j}\,\mathrm{d}x, \quad i = 0, 1, \dots, n$$

which implies

$$\int_{-1}^{1} f(x)\,\mathrm{d}x = \sum_{j=0}^{n} w_i f(x_i) + \frac{f^{(2n+2)}(\xi)}{(2n+2)!}\int_{-1}^{1}\prod_{i=0}^{n}(x-x_i)^2\,\mathrm{d}x$$

Definition 1.8 [Pivoting] Why: small $a_{kk}^{(k)}$ leads big error.

- Partial: select an element in the same column that is below the diagonal and has the largest absolute value.
- Scaled partial: first compute scale s_i for each row; then do the same as partial

Proposition 1.1

- $\|\boldsymbol{A}\|_1 = \max_{1\le j\le n}\sum_{i=1}^{m}|a_{ij}|$
- $\|\boldsymbol{A}\|_2 = \sqrt{\lambda_{\max}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})}$. If $\boldsymbol{A}$ symmetric, then $\|\boldsymbol{A}\|_2 = \rho(\boldsymbol{A}) = \max|\lambda(\boldsymbol{A})|$
- $\|\boldsymbol{A}\|_\infty = \max_{1\le i\le m}\sum_{j=1}^{n}|a_{ij}|$
- $\rho(\boldsymbol{A}) \le \|\boldsymbol{A}\|$ for any natural norm.
- $\boldsymbol{A}$ is convergent iff $\rho(\boldsymbol{A}) < 1$.

Proof. For eigen-pair $(\lambda, \boldsymbol{x})$ with $\|\boldsymbol{x}\| = 1$,

$$|\lambda| = \|\lambda\boldsymbol{x}\| \le \|\boldsymbol{A}\|\|\boldsymbol{x}\| = \|\boldsymbol{A}\|$$

■

Definition 1.9 [Jacobi's Method]

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

Or for the matrix form $\boldsymbol{x}^{(k)} = \boldsymbol{T}_j \boldsymbol{x}^{(k-1)} + \boldsymbol{c}$:

$$\boldsymbol{x}^{(k)} = \boldsymbol{D}^{-1}(\boldsymbol{L} + \boldsymbol{U})\boldsymbol{x}^{(k-1)} + \boldsymbol{D}^{-1}\boldsymbol{b},$$

where $\boldsymbol{D}$ are the diagonal of $\boldsymbol{A}$; $-\boldsymbol{L}, -\boldsymbol{U}$ are the strictly lower and upper part of $\boldsymbol{A}$. ▪

Definition 1.10 [Gauss-Seidel Method]

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

Or for the matrix form $\boldsymbol{x}^{(k)} = \boldsymbol{T}_g \boldsymbol{x}^{(k-1)} + \boldsymbol{c}_g$:

$$(\boldsymbol{D} - \boldsymbol{L})\boldsymbol{x}^{(k)} = \boldsymbol{U}\boldsymbol{x}^{(k-1)} + \boldsymbol{b} \implies \boldsymbol{x}^{(k)} = (\boldsymbol{D} - \boldsymbol{L})^{-1}\boldsymbol{U}\boldsymbol{x}^{(k-1)} + (\boldsymbol{D} - \boldsymbol{L})^{-1}\boldsymbol{b}$$

▪

Proposition 1.2 If $\rho(\boldsymbol{T}) < 1$, then $(\boldsymbol{I} - \boldsymbol{T})^{-1}$ exists, and

$$(\boldsymbol{I} - \boldsymbol{T})^{-1} = \boldsymbol{I} + \boldsymbol{T} + \boldsymbol{T}^2 + \cdots$$

Proposition 1.3 The iteration $\boldsymbol{x}^{(k)} = \boldsymbol{T}\boldsymbol{x}^{(k-1)} + \boldsymbol{c}$ converges to the unique solution to

$$\boldsymbol{x} = \boldsymbol{T}\boldsymbol{x} + \boldsymbol{c}$$

iff $\rho(\boldsymbol{T}) < 1$. The error bound holds:

- $\|\boldsymbol{x} - \boldsymbol{x}^{(k)}\| \le \|\boldsymbol{T}\|^k \|\boldsymbol{x}^{(0)} - \boldsymbol{x}\|$
- $\|\boldsymbol{x} - \boldsymbol{x}^{(k)}\| \le \frac{\|\boldsymbol{T}\|^k}{1-\|\boldsymbol{T}\|} \|\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}\|$
- $\|\boldsymbol{x} - \boldsymbol{x}^{(k)}\| \approx [\rho(\boldsymbol{T})]^k \|\boldsymbol{x}^{(0)} - \boldsymbol{x}\|$

Proof. Converse: Express $\boldsymbol{x}^{(k)}$ as

$$\boldsymbol{x}^{(k)} = \boldsymbol{T}^k \boldsymbol{x}^{(0)} + (\boldsymbol{T}^{k-1} + \cdots + \boldsymbol{T} + \boldsymbol{I})\boldsymbol{c}$$

Forward: Assume converge to $\boldsymbol{x}$, and therefore

$$\boldsymbol{x} - \boldsymbol{x}^{(k)} = \boldsymbol{T}(\boldsymbol{x} - \boldsymbol{x}^{(k-1)}) = \cdots = \boldsymbol{T}^k(\boldsymbol{x} - \boldsymbol{x}^{(0)})$$

for arbitrary $\boldsymbol{x}^{(0)}$. Taking limit implies $\lim_{k\to\infty} \boldsymbol{T}^k \boldsymbol{z} = \boldsymbol{0}$ for any $\boldsymbol{z}$. ■

Proposition 1.4 Sufficient condition for convergence of Jacobi and Gauss-Seidel method: $\boldsymbol{A}$ is strictly diagonally dominant, i.e., $|a_{ii}| > \sum_{j=1,j\neq i}^n |a_{ij}|$

Proposition 1.5 For $a_{ij} \leq 0, i \neq j$ and $a_{ii} > 0$, one and only one conditon holds:

- $0 \leq \rho(\boldsymbol{T}_g) < \rho(\boldsymbol{T}_j) < 1$
- $1 < \rho(\boldsymbol{T}_j) < \rho(\boldsymbol{T}_g)$
- $\rho(\boldsymbol{T}_j) = \rho(\boldsymbol{T}_g) = 0$
- $\rho(\boldsymbol{T}_j) = \rho(\boldsymbol{T}_g) = 1$

Definition 1.11 [Gauss-Seidel Method and Relaxation]

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

Relaxation:

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

Or equivalently, ($\omega < 1$ is under-relaxation methods, $\omega > 1$ is over-relaxation methods.)

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

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

■

Proposition 1.6 If $a_{ii} \neq 0$, then $\rho(T_\omega) \geq |\omega - 1|$, o.e., the SOR method converges only when $0 < \omega < 2$; sufficient condition for convergence: $\boldsymbol{A}$ is PD and $0 < \omega < 2$.

Proposition 1.7 If $\boldsymbol{A}$ is PD and tridiagonal, then $\rho(\boldsymbol{T}_g) = [\rho(T_j)]^2 < 1$, the optimal choice is

$$\omega = \frac{2}{1 + \sqrt{1 - [\rho(T_j)]^2}}$$

under this choice, we have $\rho(\boldsymbol{T}_\omega) = \omega - 1$

1. Direct method: Gaussian Eliminiation;

   Advantage: Exact method, no truncation error;

   Disadvantage: Computationally expansive, large round off error

   Suitable: linear systems of small dimension

2. Iterative method:

   Advantage: efficient in terms of both computer storage and computation

   Disadvantage: not so efficient for small dimension

   Suitable for: large linear sparse systems

Definition 1.12 The condition number of nonsingular matrix $\boldsymbol{A}$ is $K(\boldsymbol{A}) = \|\boldsymbol{A}\|\|\boldsymbol{A}^{-1}\|$. If close to 1, then $\boldsymbol{A}$ is well-conditioned, otherwise ill-conditioned.

Theorem 1.1 For natural norm,

$$\|\tilde{\boldsymbol{x}} - \boldsymbol{x}\| \le \|\boldsymbol{A}^{-1}\|\|\boldsymbol{r}\| = K(\boldsymbol{A})\frac{\|\boldsymbol{r}\|}{\|\boldsymbol{A}\|}$$

$$\frac{\|\tilde{\boldsymbol{x}} - \boldsymbol{x}\|}{\|\boldsymbol{x}\|} \le \|\boldsymbol{A}\|\|\boldsymbol{A}^{-1}\|\frac{\|\boldsymbol{r}\|}{\|\boldsymbol{b}\|} = K(\boldsymbol{A})\frac{\|\boldsymbol{r}\|}{\|\boldsymbol{b}\|}, \quad \boldsymbol{x} \neq 0, \boldsymbol{b} \neq 0$$

Proof. Consider the first equality and $\|\boldsymbol{b}\| \le \|\boldsymbol{A}\|\|\boldsymbol{x}\|$ ∎

Theorem 1.2 Suppose $\boldsymbol{A}$ is nonsingular and $\|\delta\boldsymbol{A}\| \le \frac{1}{\|\boldsymbol{A}\|}$, the solution $\tilde{\boldsymbol{x}}$ to $(\boldsymbol{A} + \delta\boldsymbol{A})\boldsymbol{x} = \boldsymbol{b} + \delta\boldsymbol{b}$ has the bound

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

Proposition 1.8 For $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, $\mathrm{cond}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}) = [\mathrm{cond}(\boldsymbol{A})]^2$

Theorem 1.3 For column full rank $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, we have $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$, for $\boldsymbol{Q} \in \mathbb{R}^{m\times n}$ with orthogonal columns, $\boldsymbol{R} \in \mathbb{R}^{n\times n}$ is an upper triangular matrix.

As a result, the least square solution becomes

$$\boldsymbol{x}^* = \boldsymbol{R}^{-1}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{b}$$

Gram-Schmidt Process: numerically unstable; ease of implementation

Theorem 1.4 — Householder matrix. For given vector $\boldsymbol{x}$ and unit vector $\boldsymbol{g}$, define $H := \boldsymbol{I} - 2\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}$,

$$\boldsymbol{H}\boldsymbol{x} = \|\boldsymbol{x}\|\boldsymbol{g}, \quad \boldsymbol{u} = \frac{\boldsymbol{x} - \|\boldsymbol{x}\|g}{\|\boldsymbol{x} - \|\boldsymbol{x}\|g\|}$$

$$\boldsymbol{H}_2 = \begin{pmatrix} 1 & 0 \\ 0 & \tilde{\boldsymbol{H}}_2 \end{pmatrix}$$

Define $\boldsymbol{Q} = \boldsymbol{H}_n \cdots \boldsymbol{H}_1$, and $\boldsymbol{Q}\boldsymbol{A} = \boldsymbol{R} = [\boldsymbol{R}_1; \boldsymbol{0}]$, which implies

$$\boldsymbol{A} = [\boldsymbol{Q}_1, \boldsymbol{Q}_2] \begin{pmatrix} \boldsymbol{R}_1 \\ \boldsymbol{0} \end{pmatrix} = \boldsymbol{Q}_1\boldsymbol{R}_1$$