Prelim Notes for Numerical Analysis *

Wenqiang Feng †

Abstract

This note is intended to assist my prelim examination preparation. You can download and distribute it. Please be aware, however, that the note contains typos as well as incorrect or inaccurate solutions . At here, I also would like to thank Liguo Wang for his help in some problems. This note is based on the Dr. Abner J. Salgado's lecture note [4]. Some solutions are from Dr. Steven Wise's lecture note [5].

*Key words: UTK, PDE, Prelim exam, Numerical Analysis.

†Department of Mathematics,University of Tennessee, Knoxville, TN, 37909, wfeng@math.utk..edu

Contents

List of Figures

List of Tables

1 Preliminaries

1.1 Linear Algebra Preliminaries

1.1.1 Common Properties

Properties 1.1. *(Structure of Matrices) Let $A = [A_{ij}]$ be a square or rectangular matrix, A is called*

- *diagonal : if $a_{ij} = 0$, $\forall i \neq j$,*
- *tridiagonal : if $a_{ij} = 0$, $\forall |i-j| > 1$,*
- *upper triangular : if $a_{ij} = 0$, $\forall i > j$,*
- *lower triangular : if $a_{ij} = 0$, $\forall i < j$,*
- *upper Hessenberg : if $a_{ij} = 0$, $\forall i > j+1$,*
- *lower Hessenberg : if $a_{ij} = 0$, $\forall j > i+1$,*
- *block diagonal :$A = diag(A_{11}, A_{22}, \cdots, A_{nn})$,*
- *block diagonal :$A = diag(A_{i,i-1}, A_{ii}, \cdots, A_{i,i+1})$.*

Properties 1.2. *(Type of Matrices) Let $A = [A_{ij}]$ be a square or rectangular matrix, A is called*

- *Hermitian : if $A^* = A$,*
- *skew hermitian : if $A^* = -A$,*
- *symmetric : if $A^T = A$,*
- *skew symmetric : if $A^T = -A$,*
- *normal : if $A^T A = AA^T$, when $A \in \mathbb{R}^{n\times n}$, if $A^*A = AA^*$, when $A \in \mathbb{C}^{n\times n}$,*
- *orthogonal : if $A^T A = I$, when $A \in \mathbb{R}^{n\times n}$, unitary : if $A^*A = I$, when $A \in \mathbb{C}^{n\times n}$.*

Properties 1.3. *(Properties of invertible matrices) Let A be $n \times n$ square matrix. If A is invertible , then*

- *$det(A) \neq 0$,*
- *$nullity(A) = 0$,*
- *$rank(A) = n$,*
- *$\lambda_i \neq 0$, (λ_i eigenvalues),*
- *$Ax = b$ has a unique solution for every $b \in \mathbb{R}^n$*
- *$Ax = 0$ has only trivial solution,*
- *the row vectors are linearly independent ,*
- *the column vectors are linearly independent ,*
- *the row vectors of A form a basis for $\mathbb{R}^n$.*
- *the column vectors of A form a basis for $\mathbb{R}^n$,*
- *the row vectors of A span $\mathbb{R}^n$.*
- *the column vectors of A span $\mathbb{R}^n$.*

Properties 1.4. *(Properties of conjugate transpose) Let A, B be $n \times n$ square matrix and γ be a complex constant, then*

- *$(A^*)^* = A$,*
- *$det(A^*) = det(A)$*
- *$(AB)^* = B^*A^*$,*
- *$tr(A^*) = tr(A)$*
- *$(A+B)^* = A^* + B^*$,*
- *$(\gamma A)^* = \gamma^* A^*$.*

Properties 1.5. *(Properties of similar matrices) If $A \sim B$, then*

- *$det(A) = det(B)$,*
- *$rank(A) = rank(B)$,*
- *$eig(A) = eig(B)$,*
- *if $B \sim C$, then $A \sim C$*
- *$A \sim A$,*
- *$B \sim A$*

Properties 1.6. *(Properties of Unitary Matrices) Let A be a $n \times n$ Unitary matrix, then*

- $A^* = A^{-1}$,
- A^* *is unitary,*
- A *is diagonalizable,*
- A *is unitarily similar to a diagonal matrix ,*
- *the row vectors of A form an orthonormal set,*
- $A^* = I$,
- A *is an isometry.*
- *the column vectors of A form an orthonormal set.*

Properties 1.7. *(Properties of Hermitian Matrices) Let A be a $n \times n$ Hermitian matrix, then*

- *its eigenvalues are real ,*
- A *is unitarily diagonalizable (Spectral theorem),*
- $v_i^* v_j = 0, i \neq j$, v_i, v_j *eigenvectors,*
- $A = H + K$, H *is Hermitian and* K *is skew-Hermitian,*

Properties 1.8. *(Properties of positive definite Matrices) Let $A \in \mathbb{C}^{n \times n}$ be a positive definite Matrix and $B \in \mathbb{C}^{n \times n}$, then*

- $\sigma(A) \subset (0, \infty)$,
- A *is invertible,*
- *if B is invertible, B^*B positive semidefinite ,*
- *if A is positive semidefinite then $diag(A) \geq 0$,*
- *if A is positive definite then $diag(A) > 0$.*
- B^*B *is positive semidefinite*

Properties 1.9. *(Properties of determinants) Let A, B be $n \times n$ square matrix and α be a real constant, then*

- $det(A^T) = det(A)$,
- $det(\alpha A) = \alpha^n det(A)$,
- $det(AB) = det(A)det(B)$,
- $det(A^{-1}) = \frac{1}{det(A)} = det(A)^{-1}$.

Properties 1.10. *(Properties of inverse) Let A, B be $n \times n$ square matrix and α be a real constant, then*

- $(A^*)^{-1} = (A^{-1})^*$,
- $(A^{-1})^{-1} = A$,
- $(AB)^{-1} = B^{-1}A^{-1}$,
- $(\alpha A)^{-1} = \frac{1}{\alpha} A^{-1}$
- $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, $A^{-1} = \frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$.

Properties 1.11. *(Properties of Rank) Let A be $m \times n$ matrix, B be $n \times m$ matrix and P, Q are invertible $n \times n$ matrices, then*

- $rank(A) \leq \min\{m, n\}$,
- $rank(A) = rank(A^*)$,
- $rank(A) + dim(ker(A)) = n$,
- $rank(AQ) = Rank(A) = Rank(PA)$,
- $rank(PAQ) = Rank(A)$,
- $rank(AB) \geq rank(A) + rank(B) - n$,
- $rank(AB) \leq \min\{rank(A), rank(B)\}$,
- $rank(AB) \leq rank(A) + rank(B)$.

1.1.2 Similar and diagonalization

Theorem 1.1. *(Similar) A is said to be similar to B, if there is a nonsingular matrix X, such that*

$$A = XBX^{-1}, (A \sim B).$$

Theorem 1.2. *(Diagonalizable[a]) A matrix is diagonalizable , if and only if there exist a nonsingular matrix X and a diagonal matrix D such that $A = XDX^{-1}$.*

[a] Being diagonalizable has nothing to do with being invertible.

Theorem 1.3. *(Diagonalizable) A matrix is diagonalizable , if and only if all its eigenvalues are semisimple .*

Theorem 1.4. *(Diagonalizable) Suppose $dim(A) = n$. A is said to be diagonalizable , if and only if A has n linearly independent eigenvectors .*

Corollary 1.1. *(Sample question #2, summer, 2013) Suppose $dim(A) = n$. If A has n distinct eigenvalues , then A is diagonalizable .*

Proof. (Sketch) Suppose $n = 2$, and let λ_1, λ_2 be distinct eigenvalues of A with corresponding eigenvectors v_1, v_2. Now, we will use contradiction to show v_1, v_2 are lineally independent. Suppose v_1, v_2 are lineally dependent, then

$$c_1 v_1 + c_2 v_2 = 0, \tag{1}$$

with c_1, c_2 are not both 0. Multiplying A on both sides of (210), then

$$c_1 A v_1 + c_2 A v_2 = c_1 \lambda_1 v_1 + c_2 \lambda_2 v_2 = 0. \tag{2}$$

Multiplying λ_1 on both sides of (210), then

$$c_1 \lambda_1 v_1 + c_2 \lambda_1 v_2 = 0. \tag{3}$$

Subtracting (212) form (211), then

$$c_2(\lambda_2 - \lambda_1) v_2 = 0. \tag{4}$$

Since $\lambda_1 \neq \lambda_2$ and $v_2 \neq 0$, then $c_2 = 0$. Similarly, we can get $c_1 = 0$. Hence, we get the contradiction.

A similar argument gives the result for n. Then we get A has n linearly independent eigenvectors . □

Theorem 1.5. *(Diagonalizable) Every Hermitian matrix is diagonalizable , In particular, every real symmetric matrix is diagonalizable.*

1.1.3 Eigenvalues and Eigenvectors

Theorem 1.6. *if λ is an eigenvalue of A, then $\bar{\lambda}$ is an eigenvalue of A^*.*

Theorem 1.7. *The eigenvalues of a triangular matrix are the entries on its main diagonal.*

Theorem 1.8. *Let A be square matrix with eigenvalue λ and the corresponding eigenvector x.*

- *$\lambda^n, n \in \mathbb{Z}$ is an eigenvalue of A^n with corresponding eigenvector x,*
- *if A is invertible, then $1/\lambda$ is an eigenvalue of A^{-1} with corresponding eigenvector x.*

Theorem 1.9. *Let A be $n \times n$ square matrix and let $\lambda_1, \lambda_2, \cdots, \lambda_m$ be distinct eigenvalues of A with corresponding eigenvectors $v_1, v_2, \cdots, v_m$. Then $v_1, v_2, \cdots, v_m$ are linear independent.*

1.1.4 Unitary matrices

Definition 1.1. *(Unitary Matrix) A matrix $A \in \mathbb{C}^{n\times n}$ is said to be unitary [a], if*

$$A^*A = I.$$

[a] A matrix $A \in \mathbb{R}^{n\times n}$ is said to be orthogonal , if

$$A^TA = I.$$

Theorem 1.10. *(Angle preservation) A matrix is unitary , then the transformation defined by A preserves angles.*

Proof. For any vectors $x, y \in \mathbb{C}^n$ that is angle θ is determined from the inner product via $\cos\theta = \frac{<x,y>}{\|x\|\|y\|}$. Since A is unitary (and thus an isometry), then

$$< Ax, Ay > = < A^*Ax, y > = < x, y > .$$

This proves the Angle preservation. □

Theorem 1.11. *(Angle preservation) A matrix is real orthogonal , then A has the transformation form $T(\theta)$ for some θ*

$$A = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} T(\theta) = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ \sin(\theta) & -\cos(\theta) \end{bmatrix} \tag{5}$$

Finally, we can easily establish the diagonalzableility of the unitary matrices.

Theorem 1.12. *(Shur Decomposition) A matrix $A \in \mathbb{C}^{n\times n}$ is similar to a upper triangular matrix and*

$$A = UTU^{-1}, \tag{6}$$

where U is a unitary matrix , T is an upper triangular matrix .

Proof. see Appendix (**??**) □

Theorem 1.13. *(Spectral Theorem for Unitary matrices) A is unitary , then A is diagonalizable and A is unitarily similar to a diagonal matrix .*

$$A = UDU^{-1} = UDU^*, \tag{7}$$

where U is a unitary matrix , D is an diagonal matrix .

Proof. Result follows from 1.12. □

Theorem 1.14. *(Spectral representiation) A is unitary , then*

1. *A has a set of n orthogonal eigenvectors,*
2. *let $\{v_1, v_2, \cdots, v_n\}$ be the eigenvalues w.r.t the corresponding orthogonal eigenvectors $\{\lambda_1, \lambda_2, \cdots, \lambda_n\}$. The A has the representation as the sum of rank one matrices given by*

$$A = \sum_{i=1}^{n} \lambda_i v_i v_i^T. \tag{8}$$

Note: this representation is often called the Spectral Representation or Spectral Decomposition of A.

Proof. see Appendix (**??**) □

1.1.5 Hermitian matrices

Definition 1.2. *(Hermitian Matrix) A matrix is Hermitian , if*

$$A^* = A.$$

Definition 1.3. *Let A be Hermitian , then the spectral of A, $\sigma(A)$, is real.*

Proof. Let $\lambda \in \sigma(A)$ with corresponding eigenvector v. Then

$$\begin{aligned} < Av, v > &= < \lambda v, v >= \lambda < v, v > & (9)\\ < Av, v > &= < v, A^* v >=< v, \bar{\lambda} v >= \bar{\lambda} < v, v > . & (10) \end{aligned}$$

Since $< v, v >\neq 0$,therefore $\lambda = \bar{\lambda}$. Hence λ is real. □

Definition 1.4. *Let A be Hermitian , then the different eigenvector are orthogonal i.e.*

$$< v_i, v_j >= 0, i \neq j. \tag{11}$$

Proof. Let λ_1, λ_2 be the arbitrary two different eigenvalues with corresponding eigenvector v_1, v_2. Then

$$\begin{aligned} < Av_1, v_2 > &= < \lambda_1 v_1, v_2 >= \lambda_1 < v_1, v_2 > & (12)\\ < Av_1, v_2 > &= < v_1, A^* v_2 >=< v_1, Av_2 >=< v, \lambda_2 v_2 >= \lambda_2 < v_1, v_2 > . & (13) \end{aligned}$$

Since $\lambda_1 \neq \lambda_2$,therefore $< v_1, v_2 >= 0$. □

Theorem 1.15. *(Spectral Theorem for Hermitian matrices) A is Hermitian , then A is unitary diagonalizable .*

$$A = UDU^{-1} = UDU^*, \tag{14}$$

where U is a unitary matrix , D is an diagonal matrix .

Theorem 1.16. *If A, B are unitarily similar , then A is Hermitian if and only if B is Hermitian .*

Proof. Since A, B are unitarily similar , then $A = UBU^{-1}$, where U is a unitary matrix . And

$$A^* = U^{-1^*} B^* U^* = U^{*-1} B^* U^* = UB^* U^{-1},$$

since U is a unitary matrix. Therefore

$$UBU^{-1} = A = A^* = UB^* U^{-1}.$$

Hence, $B = B^*$. □

Theorem 1.17. *If $A = A^*$, then $\rho(A) = \|A\|_2$.*

Proof. Since A is self-adjoint, there an orthonormal basis of eigenvector $x \in \mathbb{C}^n$, s.t.

$$x = \alpha_1 e_1 + \alpha_2 e_2 + \cdots + \alpha_n e_n.$$

Moreover, $Ae_i = \lambda_i e_i$, $\|e_i\| = 1$ and $(e_i, e_j) = 0$ when $i \not\equiv j$,$(e_j, e_j) = 1$. So,

$$\|x\|_{\ell^2}^2 = \sum_{i=1}^{n} |\alpha_i|^2,$$

since,

$$\begin{aligned}(x,x) &= (\sum_{i=1}^{n} \alpha_i e_i, \sum_{j=1}^{n} \alpha_j e_j) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \bar{\alpha}_j e_i e_j \\ &= \sum_{i=1}^{n} |\alpha_i|^2.\end{aligned}$$

Since, $Ax = A(\alpha_1 e_1 + \alpha_2 e_2 + \cdots + \alpha_n e_n) = \alpha_1 \lambda_1 e_1 + \alpha_2 \lambda_2 e_2 + \cdots + \alpha_n \lambda_n e_n$, then

$$\|Ax\|_{\ell^2}^2 = \sum_{i=1}^{n} |\lambda_i \alpha_i|^2 = \sum_{i=1}^{n} |\lambda_i|^2 |\alpha_i|^2 \le \max\{|\lambda_i|\}^2 \sum_{i=1}^{n} |\alpha_i|^2.$$

Therefore,

$$\|Ax\|_{\ell^2} \le \rho(A) \|x\|_{\ell^2},$$

i.e.

$$\|A\|_2 = \sup_{x \in \mathbb{C}^n} \frac{\|Ax\|_{\ell^2}}{\|x\|_{\ell^2}} \le \rho(A).$$

Let k be the index, s.t: $|\lambda_n| = \rho(A)$ and $x = e_k, Ax = Ae_k = \lambda_n e_k$, so $\|Ax\|_{\ell^2} = |\lambda_n| = \rho(A)$ and

$$\|A\|_2 = \sup_{x \in \mathbb{C}^n} \frac{\|Ax\|_{\ell^2}}{\|x\|_{\ell^2}} \ge \frac{\|Ax\|_{\ell^2}}{\|x\|_{\ell^2}} = \rho(A).$$

□

1.1.6 Positive definite matrices

Definition 1.5. *(Positive Definite Matrix)*

1. *A symmetric real matrix $A \in \mathbb{R}^{n\times n}$ is said to be Positive Definite , if*
$$x^T Ax > 0, \quad \forall x \neq 0.$$
2. *A Hermitian matrix $A \in \mathbb{C}^{n\times n}$ is said to be Positive Definite , if*
$$x^* Ax > 0, \quad \forall x \neq 0.$$

Theorem 1.18. *Let $A, B \in \mathbb{C}^{n\times n}$. Then*

1. *if A is positive definite, then $\sigma A \subset (0,\infty)$,*
2. *if A is positive definite, then A is invertible,*
3. *B^*B is positive semidefinite,*
4. *if B is invertible, then B^*B is positive definite.*
5. *if B is positive definite, then $diag(B)$ is nonnegative,*
6. *if $diag(B)$ strictly positive, thenif B is positive definite.*

Problem 1.1. *(Sample question #1, summer, 2013) Suppose $A \in \mathbb{C}^{n\times n}$ is hermitian and $\sigma(A) \subset (0,\infty)$. Prove A is Hermitian Positive Defined (HPD).*

Proof. Since, A is Hermitian, then is Unitary diagonalizable. i.e. $A = UDU^{-1} = UDU^*$, then

$$x^*Ax = x^*UDU^{-1}x = x^*UDU^*x = (U^*x)^*D(U^*x). \tag{15}$$

Moreover, since $\sigma(A) \subset (0,\infty)$ then $\tilde{x}^*D\tilde{x} > 0$ for any nonzero $\tilde{x}$. Hence

$$x^*Ax = (U^*x)^*D(U^*x) = \tilde{x}^*D\tilde{x} > 0, \text{ for any nonzero } x. \tag{16}$$

□

1.1.7 Normal matrices

Definition 1.6. *(Normal Matrix) A matrix is called normal , if*
$$A^*A = AA^*.$$

Corollary 1.2. *Unitary matrix and Hermitian matrix are normal matrices.*

Theorem 1.19. *$A \in \mathbb{C}^{n\times n}$ is normal if and only if every matrix unitarily equivalent to A is normal.*

Theorem 1.20. *$A \in \mathbb{C}^{n\times n}$ is normal if and only if every matrix unitarily equivalent to A is normal.*

Proof. Suppose A is normal and $B = U^*AU$, where U is unitary. Then $B^*B = U^*A^*UU^*AU = U^*A^*AU = U^*AA^*U = U^*AUU^*A^*U = BB^*$, so B is normal. Conversely, If B is normal, it is easy to get that $U^*A^*AU = U^*AA^*U$, then $A^*A = AA^*$ □

Theorem 1.21. *(Spectral theorem for normal matrices) If $A \in \mathbb{C}^{n\times n}$ has eigenvalues $\lambda_1, \cdots, \lambda_n$, counted according to multiplicity, the following statements are equivalent.*

1. *A is normal,*
2. *A is unitarily diagonalizable,*
3. $\sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}|^2 = \sum_{i=1}^{n}|\lambda_i|^2$,
4. *There is an orthonormal set of n eigenvectors of A.*

1.1.8 Common Theorems

Definition 1.7. *(Orthogonal Complement) Suppose $S \subset \mathbb{R}^n$ is a subspace. The (Orthogonal Complement) of S is defined as*

$$S^{\perp} = \left\{ y \in \mathbb{R}^n \mid y^T x = 0, \forall x \in S \right\}$$

Theorem 1.22. *Suppose $A \in \mathbb{R}^{n\times n}$. Then*

1. $\mathcal{R}(A)^{\perp} = \mathcal{N}(A^T)$,
2. $\mathcal{R}(A^T)^{\perp} = \mathcal{N}(A)$.

Proof. 1. For any $\tilde{y} \in \mathcal{R}(A)^{\perp}$, then $\tilde{y}^T y = 0, \forall y \in \mathcal{R}(A)$. And $\forall y \in \mathcal{R}(A)$, there exists x, such that $Ax = y$. Then

$$\tilde{y}^T A x = (A^T \tilde{y})^T x = 0.$$

Since, x is arbitrary, so it must be $A^T\tilde{y} = 0$. Hence

$$\mathcal{R}(A)^{\perp} \subset \mathcal{N}(A^T)$$

Conversely, suppose $y \in \mathcal{N}(A^T)$, then $A^T y = 0$ and hence $(A^T y)^T x = y^T A x = 0$ for any $x \in \mathbb{R}^n$. So, $y \in \mathcal{R}(A^T)^{\perp}$. Therefore

$$\mathcal{N}(A^T) \subset \mathcal{R}(A)^{\perp}$$

$\mathcal{R}(A)^{\perp} = \mathcal{N}(A^T)$,

2. Similarly, we can prove $\mathcal{R}(A^T)^{\perp} = \mathcal{N}(A)$,

□

1.2 Calculus Preliminaries

Definition 1.8. *(Taylor formula for one variable) Let $f(x)$ to be n-th differentiable at x_0, then there exists a neighborhood $B(x_0, \epsilon)$, $\forall x \in B(x_0, \epsilon)$, s.t.*

$$\begin{aligned} f(x) &= f(x_0) + f'(x_0)(x-x_0) + \frac{f''(x_0)}{2!}(x-x_0)^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n + \mathcal{O}((x-x_0)^{n+1}) \\ &= f(x_0) + f'(x_0)\Delta x + \frac{f''(x_0)}{2!}\Delta x^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}\Delta x^n + \mathcal{O}(\Delta x^{n+1}). \end{aligned} \tag{17}$$

Definition 1.9. *(Taylor formula for two variables)* *Let* $f(x,y) \in C^{k+1}(B((x_0,y_0),\epsilon))$, *then* $\forall (x_0+\Delta x, y_0+\Delta y) \in B((x_0,y_0),\epsilon))$,

$$\begin{aligned} f(x_0+\Delta x, y_0+\Delta y) &= f(x_0,y_0) + \left(\Delta x\frac{\partial}{\partial x} + \Delta y\frac{\partial}{\partial y}\right) f(x_0,y_0) \\ &+ \frac{1}{2!}\left(\Delta x\frac{\partial}{\partial x} + \Delta y\frac{\partial}{\partial y}\right)^2 f(x_0,y_0) + \cdots \\ &+ \frac{1}{k!}\left(\Delta x\frac{\partial}{\partial x} + \Delta y\frac{\partial}{\partial y}\right)^k f(x_0,y_0) + \mathcal{R}_k \end{aligned} \tag{18}$$

where

$$\mathcal{R}_k = \frac{1}{(k+1)!}\left(\Delta x\frac{\partial}{\partial x} + \Delta y\frac{\partial}{\partial y}\right)^{k+1} f(x_0+\theta\Delta x, y_0+\theta\Delta y),\ \ \theta \in (0,1).$$

Definition 1.10. *(Commonly used taylor series)*

$$\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n = 1 + x + x^2 + x^3 + x^4 + \cdots, \quad x \in (-1,1), \tag{19}$$

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots, \quad x \in \mathbb{R}, \tag{20}$$

$$\sin(x) = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)!} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!} - \cdots, \quad x \in \mathbb{R}, \tag{21}$$

$$\cos(x) = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n}}{(2n)!} = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \frac{x^8}{8!} - \cdots, \quad x \in \mathbb{R}, \tag{22}$$

$$\ln(1+x) = \sum_{n=0}^{\infty} (-1)^n \frac{x^{n+1}}{n+1} = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots, \quad x \in (-1,1). \tag{23}$$

1.3 Preliminary Inequalities

Definition 1.11. *(Cauchy's Inequality)*

$$ab \le \frac{a^2}{2} + \frac{b^2}{2}, \quad \textit{for all } a,b \in \mathbb{R}. \tag{24}$$

Proof. Since $(a-b)^2 = a^2 - 2ab + b^2 \ge 0$, therefore $ab \le \frac{a^2}{2} + \frac{b^2}{2}$, for all $a,b \in \mathbb{R}$. □

Definition 1.12. *(Cauchy's Inequality with ϵ)*

$$ab \le \epsilon a^2 + \frac{b^2}{4\epsilon}, \quad \textit{for all } a,b > 0\ ,\ \epsilon > 0. \tag{25}$$

Proof. Using Cauchy's Inequality with $\sqrt{2\epsilon}a$, $\frac{1}{\sqrt{2\epsilon}}b$ in place of a,b, we can get the result. □

Definition 1.13. *(Young's Inequality)* *Let* $1 < p,q < \infty, \frac{1}{p} + \frac{1}{q} = 1$. *Then*

$$ab \le \frac{a^p}{p} + \frac{b^q}{q} \quad \text{for all } a,b > 0\,. \tag{26}$$

Proof. Firstly, we introduce an auxiliary function

$$f(t) = \frac{t^p}{p} + \frac{1}{q} - t.$$

We know that the minimum value is at $t = 1$, since $f'(t) = t^{p-1} = 0$ at $t = 1$. Now, we setting $t = ab^{-q/p}$, we get

$$\begin{aligned} 0 \le f(ab^{-q/p}) &= \frac{(ab^{-q/p})^p}{p} + \frac{1}{q} - ab^{-q/p} \\ &= \frac{a^p b^{-q}}{p} + \frac{1}{q} - ab^{-q/p}. \end{aligned}$$

So,

$$ab^{-q/p} \le \frac{a^p b^{-q}}{p} + \frac{1}{q}.$$

Multiplying b^q on both side of the above equation yields

$$ab^{q-q/p} \le \frac{a^p}{p} + \frac{b^q}{q}.$$

Since, $\frac{1}{p} + \frac{1}{q} = 1$, so $pq = p + q$ and $q - q/p = 1$. Hence

$$ab \le \frac{a^p}{p} + \frac{b^q}{q} \quad \text{for all } a,b > 0\,.$$

□

Definition 1.14. *(Young's Inequality with ϵ)*

$$ab \le \epsilon a^p + C(\epsilon) b^q, \quad \text{for all } a,b > 0\,,\ \epsilon > 0, \tag{27}$$

Where, $C(\epsilon) = (\epsilon p)^{-p/q} q^{-1}$.

Proof. Using Young's Inequality with $(\epsilon p)^{1/p} a, \left(\frac{1}{\epsilon p}\right)^{1/p} b$ in place of a, b, we can get the result. □

Definition 1.15. *(Hölder's Inequality)* *Let* $1 < p,q < \infty, \frac{1}{p} + \frac{1}{q} = 1$. *If* $u \in L^p(U), v \in L^q(U)$, *then we have* $uv \in L^1(U)$ *and*

$$\int_U |uv| dx \le \left(\int_U |u|^p dx \right)^{1/p} \left(\int_U |v|^q dx \right)^{1/q} = \|u\|_{L^p(U)} \|v\|_{L^q(U)}\,. \tag{28}$$

Proof. Suppose $\int_U |u|^p dx \neq 0$ and $\int_U |v|^q dx \neq 0$. Otherwise, if $\int_U |u|^p dx = 0$, then $u \equiv 0$ a.e. and the Hölder's Inequality is trivial. We can use the same argument for v. Now, we define f, g as following

$$f = \frac{|u|}{\|u\|_{L^p}}, g = \frac{|v|}{\|v\|_{L^q}}. \tag{29}$$

Now applying Young's inequality for fg, we have

$$fg = \frac{|u|}{\|u\|_{L^p}} \frac{|v|}{\|v\|_{L^q}} \leq \frac{1}{p} \frac{|u|^p}{\|u\|_{L^p}^p} + \frac{1}{q} \frac{|u|^q}{\|u\|_{L^q}^q}. \tag{30}$$

Integrating it on U with respect to x, we obtain

$$\begin{aligned} \int_U \frac{|u|}{\|u\|_{L^p}} \frac{|v|}{\|v\|_{L^q}} dx &\leq \int_U \left(\frac{1}{p} \frac{|u|^p}{\|u\|_{L^p}^p} + \frac{1}{q} \frac{|u|^q}{\|u\|_{L^q}^q} \right) dx \\ &= \frac{1}{p} \frac{\int_U |u|^p dx}{\|u\|_{L^p}^p} + \frac{1}{q} \frac{\int_U |v|^p dx}{\|v\|_{L^q}^q} \\ &= \frac{1}{p} + \frac{1}{q} = 1. \end{aligned} \tag{31}$$

(31) implies that

$$\int_U |u||v| dx \leq \|u\|_{L^p} \|v\|_{L^q}. \tag{32}$$

Hence

$$\int_U |uv| dx \leq \int_U |u||v| dx \leq \|u\|_{L^p} \|v\|_{L^q}. \tag{33}$$

□

Corollary 1.3. *(Hölder's Inequality) Suppose that $u \in L^1(U), v \in L^\infty(U)$, then we have $uv \in L^1(U)$ and*

$$\int_U |uv| dx \leq \|u\|_{L^1(U)} \|v\|_{L^\infty(U)}. \tag{34}$$

Proof. Since $u \in L^1(U), v \in L^\infty(U)$, so $|uv| < \infty$ and

$$\int_U |uv| dx < \infty. \tag{35}$$

So $uv \in L^1(U)$.

$$\int_U |uv| dx \leq \int_U |u||v| dx \leq \|v\|_{L^\infty(U)} \int_U |u| dx = \|u\|_{L^1(U)} \|v\|_{L^\infty(U)}. \tag{36}$$

□

Definition 1.16. *(General Hölder's Inequality) Let $1 < p_1, \cdots, p_n < \infty$, $\frac{1}{p_1} + \cdots + \frac{1}{p_n} = \frac{1}{r}$. If $u_k \in L^{p_k}(U)$, then we have $\Pi_{k=1}^n u_i \in L^r(U)$ and*

$$\int_U |u_1 \cdots u_n|^r dx \leq \Pi_{k=1}^n \|u_i\|_{L^{p_k}}^r (U). \tag{37}$$

Proof. We will use induction to prove General Hölder's Inequality.

1. For $k = 2$, we have

$$\frac{1}{r} = \frac{1}{p_1} + \frac{1}{p_2},$$

so $r < \min(p_1, p_2)$, $L^{p_1} \subset L^r$ and $L^{p_2} \subset L^r$. Since $u_1 \in L^{p_1}$ and $u_2 \in L^{p_2}$, so $|u_1 u_2| < \infty$ and $\int_U |u_1 u_2|^r dx < \infty$. Therefore, $u_1 u_2 \in L^r(U)$.

$$1 = \frac{1}{p_1/r} + \frac{1}{p_2/r}.$$

Then applying Hölder's inequality for $|u_1 u_2|^r$, we have

$$\begin{aligned} & \int_U |u_1 u_2|^r dx \\ \leq & \left(\int_U (|u_1|^r)^{\frac{p_1}{r}} dx \right)^{\frac{r}{p_1}} \left(\int_U (|u_2|^r)^{\frac{p_2}{r}} dx \right)^{\frac{r}{p_2}} \\ = & \left(\int_U |u_1|^{p_1} dx \right)^{\frac{r}{p_1}} \left(\int_U |u_2|^{p_2} dx \right)^{\frac{r}{p_2}} \\ \leq & \|u_1\|^r_{L^{p_1}(U)} \|u_2\|^r_{L^{p_2}(U)}. \end{aligned}$$

2. Induction assumption: Assume the inequality holds for $k = n-1$, i.e. $\Pi_{k=1}^n u_i \in L^r(U)$ and

$$\int_U |u_1 \cdots u_{n-1}|^r dx \leq \Pi_{k=1}^{n-1} \|u_i\|^r_{L^{p_k}}(U).$$

3. Induction result: for $k = n$, we have

$$\frac{1}{p_1} + \cdots + \frac{1}{p_n} = \frac{1}{r}.$$

so $r < \min(p_1, p_2, \cdots, p_n)$ and $L^{p_k} \subset L^r$. Since $u_k \in L^{p_k}$, so $\Pi_{k=1}^n |u_i| \in L^r(U) < \infty$ and $\int_U |u_1 \cdots u_n|^r dx < \infty$. Therefore, $\Pi_{k=1}^n u_i \in L^r(U)$. let

$$\frac{1}{p_1} + \cdots + \frac{1}{p_{n-1}} = \frac{1}{p}.$$

so

$$\frac{1}{p} + \frac{1}{p_n} = \frac{1}{r}.$$

From the Hölder's inequality for $n = 2$ and the induction assumption, we have

$$\begin{aligned} \int_U |u_1 \cdots u_n|^r dx & \leq \left(\int_U |u_1 \cdots u_{n-1}|^p dx \right)^{\frac{r}{p}} \left(\int_U |u_n|^{p_n} dx \right)^{\frac{r}{p_n}} \\ & \leq \|u_1\|^r_{L^p(U)} \|u_2\|^r_{L^{p_n}(U)} = \Pi_{k=1}^n \|u_i\|^r_{L^{p_k}}(U). \end{aligned}$$

□

Corollary 1.4. *(General Hölder's Inequality) Let* $1 < p_1, \cdots, p_n < \infty, \frac{1}{p_1} + \cdots + \frac{1}{p_n} = 1$. *If* $u_k \in L^{p_k}(U)$, *then we have* $\Pi_{k=1}^n u_i \in L^1(U)$ *and*

$$\int_U |u_1 \cdots u_n| dx \leq \Pi_{k=1}^n \|u_i\|_{L^{p_k}}(U). \tag{38}$$

for $k = 1, 2, \cdots, n-1$.

Proof. Take $r = 1$ in last General Hölder's Inequality. □

Definition 1.17. *(Discrete Hölder's Inequality) Let* $1 < p, q < \infty, \frac{1}{p} + \frac{1}{q} = 1$. *Then for all* $a_k, b_k \in \mathbb{R}^n$,

$$\left| \sum_{k=1}^n a_k b_k \right| \leq \left(\sum_{k=1}^n |a_k|^p \right)^{1/p} \left(\sum_{k=1}^n |b_k|^q \right)^{1/q}. \tag{39}$$

Proof. The idea of proof is same to the integral version. Suppose $\sum |a_k|^p \neq 0$ and $\sum |b_k|^q \neq 0$. Otherwise, if $\sum |a_k|^p = 0$, then $a_k \equiv 0$ and the Hölder's Inequality is trivial. We can use the same argument for b_k. Now, we define f, g as following

$$f_k = \frac{a_k}{\|a\|_{\ell^p}}, g_k = \frac{b_k}{\|b\|_{\ell^q}}. \tag{40}$$

Now applying Young's inequality for fg, we have

$$f_k g_k = \frac{a_k}{\|a\|_{\ell^p}} \frac{b_k}{\|b\|_{\ell^q}} \leq \frac{1}{p} \frac{a_k^p}{\|a\|_{\ell^p}^p} + \frac{1}{q} \frac{b_k^q}{\|b\|_{\ell^q}^q}. \tag{41}$$

Taking summation yields

$$\begin{aligned} \sum_{k=1}^\infty f_k g_k &= \sum_{k=1}^\infty \frac{a_k}{\|a\|_{\ell^p}} \frac{b_k}{\|b\|_{\ell^q}} \\ &\leq \sum_{k=1}^\infty \left(\frac{1}{p} \frac{a_k^p}{\|a\|_{\ell^p}^p} + \frac{1}{q} \frac{b_k^q}{\|b\|_{\ell^q}^q} \right) \\ &= \frac{1}{p} \frac{\sum_{k=1}^\infty a_k^p}{\|a\|_{\ell^p}^p} + \frac{1}{q} \frac{\sum_{k=1}^\infty b_k^q}{\|b\|_{\ell^q}^q} \\ &= \frac{1}{p} + \frac{1}{q} = 1. \end{aligned} \tag{42}$$

Therefore

$$\sum_{k=1}^\infty a_k b_k \leq \|a\|_{\ell^p} \|b\|_{\ell^q}. \tag{43}$$

□

Corollary 1.5. *(Discrete Hölder's Inequality) Let* $a_k \in \ell^1$ *and* $b_k \in \ell^\infty$. *Then* $a_k b_k \in \ell^1$ *and*

$$\left| \sum_{k=1}^n a_k b_k \right| \leq \left(\sum_{k=1}^n |a_k| \right) \left(\sup_{k \in \mathbb{N}} |b_k| \right). \tag{44}$$

Proof.

$$\left|\sum_{k=1}^{n} a_k b_k\right| \leq \sum_{k=1}^{n} |a_k b_k| \leq \sum_{k=1}^{n} |a_k||b_k| \leq \left(\sum_{k=1}^{n} \sup_{k\in\mathbb{N}}(|b_k|)|a_k|\right) \leq \left(\sum_{k=1}^{n} |a_k|\right)\left(\sup_{k\in\mathbb{N}}|b_k|\right). \tag{45}$$

□

Definition 1.18. *(Cauchy-Schwarz's Inequality) Let $u, v \in L^2(U)$. Then*

$$|uv|^2 \leq \|u\|_{L^2(U)} \|v\|_{L^2(U)}. \tag{46}$$

Proof. Take $p = q = 2$ in Hölder's inequality. □

Definition 1.19. *(Discrete Cauchy-Schwarz's Inequality)*

$$\left|\sum_{i=1}^{n} x_i y_i\right|^2 \leq \sum_{i=1}^{n} |x_i|^2 \sum_{i=1}^{n} |y_i|^2. \tag{47}$$

Proof. Take $p = q = 2$ in Discrete Hölder's inequality. □

Definition 1.20. *(Minkowski's Inequality) Let $1 \leq p < \infty$ and $u, v \in L^p(U)$. Then*

$$\|u + v\|_{L^p(U)} \leq \|u\|_{L^p(U)} + \|v\|_{L^p(U)}. \tag{48}$$

Proof. Suppose $\int_U |u|^p dx \neq 0$ and $\int_U |v|^q dx \neq 0$. Otherwise, if $\int_U |u|^p dx = 0$, then $u \equiv 0$ a.e.. We can use the same argument for v. Then the Minkowski's Inequality is trivial. First, We have the following fact

$$|u + v|^p \leq (|u| + |v|)^p \leq 2^p \max(|u|^p, |v|^p) \leq 2^p (|u|^p + |v|^p) < \infty. \tag{49}$$

Hence $u + v \in L^p(U)$ if $u, v \in L^p(U)$. Let

$$\frac{1}{p} + \frac{1}{q} = 1 \text{ or } q = \frac{p}{p-1}. \tag{50}$$

Then, we have the fact that if $u + v \in L^p$ then $|u + v|^{p-1} \in L^q$, since $|u + v|^{p-1} < \infty$ and

$$\left\||u + v|^{p-1}\right\|_{L^q} = \left(\int_U \left(|u + v|^{p-1}\right)^q dx\right)^{\frac{1}{q}} = \left(\int_U |u + v|^p dx\right)^{\frac{1}{p}\cdot(p-1)} = \|u + v\|_{L^p}^{p-1} < \infty. \tag{51}$$

Now, we can use Hölder's inequality for $|u + v| \cdot |u + v|^{p-1}$, i.e.

$$\begin{aligned}
\|u + v\|_{L^p}^p = \int_U |u + v|^p dx &= \int_U |u + v||u + v|^{p-1} dx \\
&\leq \int_U |u||u + v|^{p-1} + |v||u + v|^{p-1} dx \\
&\leq \int_U |u||u + v|^{p-1} dx + \int_U |v||u + v|^{p-1} dx \\
&\leq \|u\|_{L^p} \left\||u + v|^{p-1}\right\|_{L^q} + \|v\|_{L^p} \left\||u + v|^{p-1}\right\|_{L^q} \\
&= \left(\|u\|_{L^p} + \|v\|_{L^p}\right) \left\||u + v|^{p-1}\right\|_{L^q} \\
&= \left(\|u\|_{L^p} + \|v\|_{L^p}\right) \|u + v\|_{L^p}^{p-1}.
\end{aligned} \tag{52}$$

Since $\|u+v\|_{L^p} \neq 0$, dividing $\|u+v\|_{L^p}^{p-1}$ on both side of (52) yields

$$\|u+v\|_{L^p} \leq \|u\|_{L^p} + \|v\|_{L^p}. \tag{53}$$

□

Definition 1.21. *(Discrete Minkowski's Inequality) Let $1 \leq p < \infty$ and $a_k, b_k \in L^p(U)$. Then $u+v \in L^p(U)$ and*

$$\left(\sum_{k=1}^{n} |a_k + b_k|^p\right)^{1/p} \leq \left(\sum_{k=1}^{n} |a_k|^p\right)^{1/p} + \left(\sum_{k=1}^{n} |b_k|^p\right)^{1/p}. \tag{54}$$

Proof. The idea is similar to the continuous case.

$$\begin{aligned}
\sum_{k=1}^{n} |a_k + b_k|^p &= \sum_{k=1}^{n} |a_k + b_k||a_k + b_k|^{p-1} \\
&\leq \sum_{k=1}^{n} |a_k||a_k + b_k|^{p-1} + |b_k||a_k + b_k|^{p-1} \\
&\leq \left(\sum_{k=1}^{n} |a_k|^p\right)^{1/p} \left(\sum_{k=1}^{n} \left[|a_k + b_k|^{p-1}\right]^q\right)^{1/q} \\
&+ \left(\sum_{k=1}^{n} |b_k|^p\right)^{1/p} \left(\sum_{k=1}^{n} \left[|a_k + b_k|^{p-1}\right]^q\right)^{1/q} \quad \left(\frac{1}{p} + \frac{1}{q} = 1\right) \\
&= \left(\sum_{k=1}^{n} |a_k|^p\right)^{1/p} \left(\sum_{k=1}^{n} |a_k + b_k|^p\right)^{1/q} \\
&+ \left(\sum_{k=1}^{n} |b_k|^p\right)^{1/p} \left(\sum_{k=1}^{n} |a_k + b_k|^p\right)^{1/q} \\
&= \left(\sum_{k=1}^{n} |a_k|^p\right)^{1/p} \left(\sum_{k=1}^{n} |a_k + b_k|^p\right)^{\frac{p-1}{p}} \\
&+ \left(\sum_{k=1}^{n} |b_k|^p\right)^{1/p} \left(\sum_{k=1}^{n} |a_k + b_k|^p\right)^{\frac{p-1}{p}} \\
&= \left(\left(\sum_{k=1}^{n} |a_k|^p\right)^{1/p} + \left(\sum_{k=1}^{n} |b_k|^p\right)^{1/p}\right)\left(\sum_{k=1}^{n} |a_k + b_k|^p\right)^{\frac{p-1}{p}}.
\end{aligned}$$

Diving $\left(\sum_{k=1}^{n} |a_k + b_k|^p\right)^{1-\frac{1}{p}}$ on both sides of the above equation, we get

$$\left(\sum_{k=1}^{n} |a_k + b_k|^p\right)^{1/p} \leq \left(\sum_{k=1}^{n} |a_k|^p\right)^{1/p} + \left(\sum_{k=1}^{n} |b_k|^p\right)^{1/p}.$$

□

Definition 1.22. *(Integral Minkowski's Inequality)* *Let $1 \le p < \infty$ and $u(x,y) \in L^p(U)$. Then*

$$\left(\int \left| \int u(x,y)dx \right|^p dy \right)^{\frac{1}{p}} \le \int \left(\int |u(x,y)|^p dy \right)^{\frac{1}{p}} dx. \tag{55}$$

Proof. 1. When $p = 1$, then

$$\int \left| \int u(x,y)dx \right| dy \le \int \int \left| u(x,y) \right| dxdy = \int \int \left| u(x,y) \right| dydx. \tag{56}$$

Where, the last step follows by Fubini's theorem for nonnegative measurable functions.

2. When $1 < p < \infty$,

$$\begin{aligned}
& \int \left| \int u(x,y)dx \right|^p dy \\
\le\ & \int \left(\int \left| u(x,y) \right| dx \right)^p dy \\
=\ & \int \underbrace{\left(\int \left| u(x,y) \right| dx \right)^{p-1}}_{\text{independent on x}} \left(\int \left| u(x,y) \right| dx \right) dy \\
=\ & \int \int \left(\int \left| u(x,y) \right| dx \right)^{p-1} \left| u(x,y) \right| dxdy \\
=\ & \int \int \left(\int \left| u(x,y) \right| dx \right)^{p-1} \left| u(x,y) \right| dydx \ \text{(Fubini)} \\
=\ & \int \int \left(\int \left| u(x,y) \right| dx \right)^{p-1} \left| u(x,y) \right| dydx \\
\le\ & \int \left(\int \left(\int \left| u(x,y) \right| dx \right)^{(p-1)q} dy \right)^{1/q} \left(\int \left| u(x,y) \right|^p dy \right)^{1/p} dx \ \text{(Hölder's)} \\
=\ & \int \left(\underbrace{\int \left(\int \left| u(x,y) \right| dx \right)^p dy}_{\text{constant}} \right)^{1/q} \left(\int \left| u(x,y) \right|^p dy \right)^{1/p} dx \quad (\frac{1}{p} + \frac{1}{q} = 1) \\
=\ & \left(\int \left(\int \left| u(x,y) \right| dx \right)^p dy \right)^{1/q} \int \left(\int \left| u(x,y) \right|^p dy \right)^{1/p} dx
\end{aligned}$$

So, we get

$$\int \left(\int \left| u(x,y) \right| dx \right)^p dy \le \left(\int \left(\int \left| u(x,y) \right| dx \right)^p dy \right)^{1-1/p} \int \left(\int \left| u(x,y) \right|^p dy \right)^{1/p} dx.$$

dividing $\left(\int \left(\int \left| u(x,y) \right| dx \right)^p dy \right)^{1-1/p}$ on both sides of the above equation yields

$$\left(\int \left(\int \left| u(x,y) \right| dx \right)^p dy \right)^{1/p} \le \int \left(\int \left| u(x,y) \right|^p dy \right)^{1/p} dx.$$

Hence, we proved the result by the following fact

$$\left(\int\left|\int u(x,y)dx\right|^p dy\right)^{1/p} \le \left(\int\left(\int |u(x,y)|dx\right)^p dy\right)^{1/p}.$$

□

Definition 1.23. *(Differential Version of Gronwall's Inequality) Let $\eta(\cdot)$ be a nonnegative, absolutely continuous function on $[0,T]$, which satisfies for a.e t the differential inequality*

$$\eta'(t) \le \phi(t)\eta(t) + \psi(t), \tag{57}$$

where $\phi(t)$ and $\psi(t)$ are nonnegative, summable functions on $[0,T]$. Then

$$\eta(t) \le e^{\int_0^t \phi(s)ds}\left[\eta(0) + \int_0^t \psi(s)ds\right], \forall 0 \le t \le T. \tag{58}$$

In particular, if

$$\eta' \le \phi\eta, \text{ on} [0,T] \text{ and } \eta(0) = 0, \tag{59}$$

$$\eta(t) = 0, \forall 0 \le t \le T. \tag{60}$$

Proof. Since

$$\eta'(t) \le \phi(t)\eta(t) + \psi(t), a.e. 0 \le t \le T. \tag{61}$$

then

$$\eta'(s) - \phi(s)\eta(s) \le \psi(s), a.e. 0 \le s \le T. \tag{62}$$

Let

$$f(s) = \eta(s)e^{-\int_0^s \phi(\xi)d\xi}. \tag{63}$$

By product rule and chain rule, we have

$$\begin{aligned}
\frac{df}{ds} &= \eta'(s)e^{-\int_0^s \phi(\xi)d\xi} - \eta(s)e^{-\int_0^s \phi(\xi)d\xi}\phi(s), &(64)\\
&= (\eta'(s) - \eta(s)\phi(s))e^{-\int_0^s \phi(\xi)d\xi} &(65)\\
&\le \psi(s)e^{-\int_0^s \phi(\xi)d\xi}, a.e. 0 \le t \le T. &(66)
\end{aligned}$$

Integral the above equation from 0 to t, then we get

$$\int_0^t \eta(s)e^{-\int_0^s \phi(\xi)d\xi}ds = \eta(t)e^{-\int_0^t \phi(\xi)d\xi} - \eta(0) \le \int_0^t \psi(s)e^{-\int_0^s \phi(\xi)d\xi}ds,$$

i.e.

$$\eta(t)e^{-\int_0^t \phi(\xi)d\xi} \le \eta(0) + \int_0^t \psi(s)e^{-\int_0^s \phi(\xi)d\xi}ds.$$

Therefore

$$\eta(t) \le e^{\int_0^t \phi(\xi)d\xi}\left[\eta(0) + \int_0^t \psi(s)e^{-\int_0^s \phi(\xi)d\xi}ds\right].$$

□

Definition 1.24. *(Integral Version of Gronwall's Inequality) Let $\xi(\cdot)$ be a nonnegative, summable function on $[0, T]$, which satisfies for a.e t the integral inequality*

$$\xi(t) \le C_1 \int_0^t \xi(s)ds + C_2, \tag{67}$$

where $C_1, C_2 \ge 0$. Then

$$\xi(t) \le C_2\left(1 + C_1 t e^{C_1 t}\right), \ \forall a.e.\ 0 \le t \le T. \tag{68}$$

In particular, if

$$\xi(t) \le C_1 \int_0^t \xi(s)ds, \ \forall a.e.\ 0 \le t \le T, \tag{69}$$

$$\xi(t) = 0, a.e. \tag{70}$$

Proof. Let

$$\eta(t) := \int_0^t \xi(s)ds, \tag{71}$$

then

$$\eta'(t) = \xi(t). \tag{72}$$

Since

$$\xi(t) \le C_1 \int_0^t \xi(s)ds + C_2, \tag{73}$$

so

$$\eta'(t) \le C_1\eta(t) + C_2. \tag{74}$$

By Differential Version of Gronwall's Inequality, we get

$$\eta(t) \le e^{\int_0^t C_1 ds}[\eta(0) + \int_0^t C_2 ds], \tag{75}$$

i.e.

$$\eta(t) \le C_2 t e^{C_1 t}. \tag{76}$$

Therefore

$$\int_0^t \xi(s)ds \le C_2 t e^{C_1 t}. \tag{77}$$

Taking derivative w.r.t t on both side of the above, we get

$$\xi(t) \le C_2 e^{C_1 t} + C_2 t e^{C_1 t} C_1 = C_2(1 + C_1 t e^{C_1 t}). \tag{78}$$

□

Definition 1.25. *(Discrete Version of Gronwall's Inequality) If*

$$(1+\gamma)a_{n+1} \le a_n + \beta f_n,\ \beta,\gamma \in \mathbb{R}, \gamma \neq -1, n = 0,\cdots, \tag{79}$$

then,

$$a_{n+1} \le \frac{a_0}{(1+\gamma)^{n+1}} + \beta \sum_{k=0}^{n} \frac{f_k}{(1+\gamma)^{n-k+1}}. \tag{80}$$

Proof. We will use induction to prove this discrete Gronwall's inequality.

1. For $n = 0$, then

$$(1+\gamma)a_1 \le a_0 + \beta f_0, \tag{81}$$

so

$$a_1 \le \frac{a_0}{(1+\gamma)} + \beta \frac{f_0}{(1+\gamma)^{n-k}}. \tag{82}$$

2. Induction Assumption: Assume the discrete Gronwall's inequality is valid for $k = n-1$, i.e.

$$a_n \le \frac{a_0}{(1+\gamma)^{n}} + \beta \sum_{k=0}^{n-1} \frac{f_k}{(1+\gamma)^{n-k}}. \tag{83}$$

3. Induction Result: For $k = n$, we have

$$\begin{aligned}
(1+\gamma)a_{n+1} &\le a_n + \beta f_n \\
&\le \frac{a_0}{(1+\gamma)^{n}} + \beta \sum_{k=0}^{n-1} \frac{f_k}{(1+\gamma)^{n-k}} + \beta f_n \\
&\le \frac{a_0}{(1+\gamma)^{n}} + \beta \sum_{k=0}^{n-1} \frac{f_k}{(1+\gamma)^{n-k}} + \beta \frac{f_n}{(1+\gamma)^{n-n}} \\
&= \frac{a_0}{(1+\gamma)^{n}} + \beta \sum_{k=0}^{n} \frac{f_k}{(1+\gamma)^{n-k}}.
\end{aligned} \tag{84}$$

Dividing $1+\gamma$ on both sides of the above equation gives

$$a_{n+1} \le \frac{a_0}{(1+\gamma)^{n+1}} + \beta \sum_{k=0}^{n} \frac{f_k}{(1+\gamma)^{n-k+1}}. \tag{85}$$

□

Definition 1.26. *(Interpolation Inequality for L^p-norm) Assume $1 \le p \le r \le q \le \infty$ and*

$$\frac{1}{r} = \frac{\theta}{p} + \frac{1-\theta}{q}. \tag{86}$$

Suppose also $u \in L^p(U) \cap L^q(U)$. Then $u \in L^r(U)$, and

$$\|u\|_{L^r(U)} \le \|u\|^{\theta}_{L^p(U)} \|u\|^{1-\theta}_{L^q(U)}. \tag{87}$$

Proof. If $1 \le p < r < q$ then $\frac{1}{q} < \frac{1}{r} < \frac{1}{p}$, hence there exists $\theta \in [0,1]$ s.t. $\frac{1}{r} = \theta\frac{1}{p} + (1-\theta)\frac{1}{q}$, therefore:

$$1 = \frac{r\theta}{p} + \frac{r(1-\theta)}{q} = \frac{1}{\frac{p}{r\theta}} + \frac{1}{\frac{q}{r(1-\theta)}}. \tag{88}$$

And $|u|^{r\theta} \in L^{\frac{p}{r\theta}}$, $|u|^{r(1-\theta)} \in L^{\frac{q}{r(1-\theta)}}$, since

$$\left(\int_U \left(|u|^{r\theta}\right)^{\frac{p}{r\theta}} dx\right)^{\frac{r\theta}{p}} = \left(\int_U |u|^p dx\right)^{\frac{r\theta}{p}} = \|u\|^{r\theta}_{L^p(U)} < \infty, \tag{89}$$

$$\left(\int_U \left(|u|^{r(1-\theta)}\right)^{\frac{q}{r(1-\theta)}} dx\right)^{\frac{r(1-\theta)}{q}} = \left(\int_U |u|^q dx\right)^{\frac{r(1-\theta)}{q}} = \|u\|^{r(1-\theta)}_{L^q(U)} < \infty. \tag{90}$$

Now, we can use Hölder's inequality for $|u|^r = |u|^{r\theta}|u|^{r(1-\theta)}$, i.e.

$$\begin{aligned}
\int_U |u|^r dx &= \int_U |u|^{r\theta}|u|^{r(1-\theta)} dx \\
&\le \left(\int_U \left(|u|^{r\theta}\right)^{\frac{p}{r\theta}} dx\right)^{\frac{r\theta}{p}} \left(\int_U \left(|u|^{r(1-\theta)}\right)^{\frac{q}{r(1-\theta)}} dx\right)^{\frac{r(1-\theta)}{q}}. &(91)\\
&= \|u\|^{r\theta}_{L^p(U)} \|u\|^{r(1-\theta)}_{L^q(U)}. &(92)
\end{aligned}$$

Therefore

$$\|u\|_{L^r(U)} \le \|u\|^{\theta}_{L^p(U)} \|u\|^{1-\theta}_{L^q(U)}. \tag{93}$$

□

Definition 1.27. *(Interpolation Inequality for L^p-norm) Assume $1 \le p \le r \le q \le \infty$ and $f \in L^q$. Suppose also $u \in L^p(U) \cap L^q(U)$. Then $u \in L^r(U)$,*

$$\|u\|_{L^r(U)} \le \|u\|^{\frac{1/p-1/r}{1/p-1/q}}_{L^p(U)} \|u\|^{\frac{1/r-1/q}{1/p-1/q}}_{L^q(U)}. \tag{94}$$

Proof. If $1 \le p < r < q$ then $\frac{1}{q} < \frac{1}{r} < \frac{1}{p}$, hence there exists $\theta \in [0,1]$ s.t. $\frac{1}{r} = \theta\frac{1}{p} + (1-\theta)\frac{1}{q}$, therefore:

$$1 = \frac{r\theta}{p} + \frac{r(1-\theta)}{q} = \frac{1}{\frac{p}{r\theta}} + \frac{1}{\frac{q}{r(1-\theta)}}. \tag{95}$$

And $|u|^{r\theta} \in L^{\frac{p}{r\theta}}$, $|u|^{r(1-\theta)} \in L^{\frac{q}{r(1-\theta)}}$, since

$$\left(\int_U \left(|u|^{r\theta}\right)^{\frac{p}{r\theta}} dx\right)^{\frac{r\theta}{p}} = \left(\int_U |u|^p dx\right)^{\frac{r\theta}{p}} = \|u\|^{r\theta}_{L^p(U)} < \infty, \tag{96}$$

$$\left(\int_U \left(|u|^{r(1-\theta)}\right)^{\frac{q}{r(1-\theta)}} dx\right)^{\frac{r(1-\theta)}{q}} = \left(\int_U |u|^q dx\right)^{\frac{r(1-\theta)}{q}} = \|u\|^{r(1-\theta)}_{L^q(U)} < \infty. \tag{97}$$

Now, we can use Hölder's inequality for $|u|^r = |u|^{r\theta}|u|^{r(1-\theta)}$, i.e.

$$\begin{aligned}\int_U |u|^r dx &= \int_U |u|^{r\theta}|u|^{r(1-\theta)} dx \\ &\leq \left(\int_U \left(|u|^{r\theta}\right)^{\frac{p}{r\theta}} dx\right)^{\frac{r\theta}{p}} \left(\int_U \left(|u|^{r(1-\theta)}\right)^{\frac{q}{r(1-\theta)}} dx\right)^{\frac{r(1-\theta)}{q}}. \quad (98)\\ &= \|u\|^{r\theta}_{L^p(U)} \|u\|^{r(1-\theta)}_{L^q(U)}. \quad (99)\end{aligned}$$

Therefore

$$\|u\|_{L^r(U)} \leq \|u\|^{\theta}_{L^p(U)} \|u\|^{1-\theta}_{L^q(U)}. \quad (100)$$

Let $\theta = \frac{1/p-1/r}{1/p-1/q}$, then we get

$$\|u\|_{L^r(U)} \leq \|u\|^{\frac{1/p-1/r}{1/p-1/q}}_{L^p(U)} \|u\|^{\frac{1/r-1/q}{1/p-1/q}}_{L^q(U)}. \quad (101)$$

□

Theorem 1.23. *(1D Dirichlet-Poincaré inequality) Let $a > 0$, $u \in C^1([-a,a])$ and $u(-a) = 0$, then the 1D Dirichlet-Poincaré inequality is defined as follows*

$$\int_{-a}^{a} \left|u(x)\right|^2 dx \leq 4a^2 \int_{-a}^{a} \left|u'(x)\right|^2 dx.$$

Proof. Since $u(-a) = 0$, then by calculus fact, we have

$$u(x) = u(x) - u(-a) = \int_{-a}^{x} u'(\xi) d\xi.$$

Therefore

$$\begin{aligned}\left|u(x)\right| &\leq \left|\int_{-a}^{x} u'(\xi) d\xi\right| \\ &\leq \int_{-a}^{x} \left|u'(\xi)\right| d\xi \\ &\leq \int_{-a}^{a} \left|u'(\xi)\right| d\xi \ (x \leq a) \\ &\leq \left(\int_{-a}^{a} 1^2 d\xi\right)^{1/2} \left(\int_{-a}^{a} \left|u'(\xi)\right|^2 d\xi\right)^{1/2} \text{ (Cauchy-Schwarz inequality)} \\ &= (2a)^{1/2} \left(\int_{-a}^{a} \left|u'(\xi)\right|^2 d\xi\right)^{1/2}.\end{aligned}$$

Therefore

$$\left|u(x)\right|^2 \leq 2a \int_{-a}^{a} \left|u'(\xi)\right|^2 d\xi.$$

Integration on both sides of the above equation from $-a$ to a w.r.t. x yields

$$\begin{aligned}\int_{-a}^{a}\left|u(x)\right|^2 dx &\leq \int_{-a}^{a} 2a\int_{-a}^{a}\left|u'(\xi)\right|^2 d\xi dx\\ &= \int_{-a}^{a}\left|u'(\xi)\right|^2 d\xi\int_{-a}^{a} 2a dx\\ &= 4a^2\int_{-a}^{a}\left|u'(\xi)\right|^2 d\xi\\ &= 4a^2\int_{-a}^{a}\left|u'(x)\right|^2 dx.\end{aligned}$$

□

Theorem 1.24. *(1D Neumann-Poincaré inequality) Let $a>0$, $u\in C^1([-a,a])$ and $\bar{u}=\fint_{-a}^{a}u(x)dx$, then the 1D Neumann-Poincaré inequality is defined as follows*

$$\int_{-a}^{a}\left|u(x)-\bar{u}(x)\right|^2 dx \leq 2a(a-c)\int_{-a}^{a}\left|u'(x)\right|^2 dx.$$

Proof. Since, $\bar{u}=\fint_{-a}^{a}u(x)dx$, then by intermediate value theorem, there exists a $c\in[-a,a]$, s.t.

$$u(c)=\bar{u}(x).$$

then by calculus fact, we have

$$u(x)-\bar{u}(x)=u(x)-u(c)=\int_c^x u'(\xi)d\xi.$$

Therefore

$$\begin{aligned}\left|u(x)-\bar{u}(x)\right| &\leq \left|\int_c^x u'(\xi)d\xi\right|\\ &\leq \int_c^x\left|u'(\xi)\right|d\xi\\ &\leq \int_c^a\left|u'(\xi)\right|d\xi \ (x\leq a)\\ &\leq \left(\int_c^a 1^2 d\xi\right)^{1/2}\left(\int_c^a\left|u'(\xi)\right|^2 d\xi\right)^{1/2} \ \text{(Cauchy-Schwarz inequality)}\\ &= (a-c)^{1/2}\left(\int_{-a}^{a}\left|u'(\xi)\right|^2 d\xi\right)^{1/2}.\end{aligned}$$

Therefore

$$\left|u(x)-\bar{u}(x)\right|^2 \leq (a-c)\int_{-a}^{a}\left|u'(\xi)\right|^2 d\xi.$$

Integration on both sides of the above equation from $-a$ to a w.r.t. x yields

$$\begin{aligned}\int_{-a}^{a}\left|u(x)-\bar{u}(x)\right|^2 dx &\leq \int_{-a}^{a}(a-c)\int_{-a}^{a}\left|u'(\xi)\right|^2 d\xi dx\\ &= \int_{-a}^{a}\left|u'(\xi)\right|^2 d\xi \int_{-a}^{a}(a-c)dx\\ &= 2a(a-c)\int_{-a}^{a}\left|u'(\xi)\right|^2 d\xi\\ &= 2a(a-c)\int_{-a}^{a}\left|u'(x)\right|^2 dx.\end{aligned}$$

□

1.4 Norms' Preliminaries

1.4.1 Vector Norms

Definition 1.28. *(Vector Norms) A vector norm is a function $\|\cdot\|:\mathbb{R}^n\mapsto\mathbb{R}$ satisfying the following conditions for all $x,y\in\mathbb{R}^n$ and $\alpha\in\mathbb{R}$*

1. *nonnegative :* $\|x\|\geq 0,\ (\|x\|=0\Leftrightarrow x=0)$,
2. *homegenity :* $\|\alpha x\|=|\alpha|\|x\|$,
3. *triangle inequality :* $\|x+y\|\leq\|x\|+\|y\|,\ \forall x,y\in\mathbb{R}^n$,

Definition 1.29. *For $x\in\mathbb{R}^n$, some of the most frequently used vector norms are*

1. *1-norm :* $\|x\|_1=\sum_{i=1}^{n}|x_i|$,
2. *2-norm :* $\|x\|_2=\sqrt{\sum_{i=1}^{n}|x_i|^2}$,
3. *∞-norm :* $\|x\|_\infty=\max_{1\leq i\leq n}|x_i|$,
4. *p-norm :* $\|x\|_p=\left(\sum_{i=1}^{n}|x_i|^p\right)^{1/p}$.

Corollary 1.6. *For all $x\in\mathbb{R}^n$,*

$$\|x\|_2\leq\|x\|_1\leq\sqrt{n}\|x\|_2, \tag{102}$$
$$\|x\|_\infty\leq\|x\|_2\leq\sqrt{n}\|x\|_\infty, \tag{103}$$
$$\frac{1}{\sqrt{n}}\|x\|_1\leq\|x\|_2\leq\sqrt{n}\|x\|_1, \tag{104}$$
$$\|x\|_\infty\leq\|x\|_1\leq\sqrt{n}\|x\|_\infty. \tag{105}$$

Theorem 1.25. *(vector 2-norm invariance) Vector 2-norm is invariant under the orthogonal transformation, i.e., if Q is an $n\times n$ orthogonal matrix, then*

$$\|Qx\|_2=\|x\|_2,\ \forall x\in\mathbb{R}^n \tag{106}$$

Proof.

$$\|Qx\|_2^2=(Qx)^T(Qx)=x^TQ^TQx=x^Tx=\|x\|_2^2.$$

□

1.4.2 Matrix Norms

Definition 1.30. *(Matrix Norms) A matrix norm is a function $\|\cdot\| : \mathbb{R}^{m\times n} \vdash \mathbb{R}$ satisfying the following conditions for all $A, B \in \mathbb{R}^{m\times n}$ and $\alpha \in \mathbb{R}$*

1. *nonnegative :* $\|x\| \geq 0, \ (\|x\| = 0 \Leftrightarrow x = 0)$,
2. *homegenity :* $\|\alpha x\| = |\alpha|\|x\|$,
3. *triangle inequality :* $\|x + y\| \leq \|x\| + \|y\|, \ \forall x, y \in \mathbb{R}^n$,

Definition 1.31. *For $A \in \mathbb{R}^{m\times n}$, some of the most frequently matrix vector norms are*

1. *F-norm :* $\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{i=1}^{n}|a_{ij}|^2}$,
2. *1-norm :* $\|A\|_1 = \max_{1\leq j\leq n}\sum_{i=1}^{m}|a_{ij}|$,
3. *∞-norm :* $\|A\|_\infty = \max_{1\leq i\leq m}\sum_{j=1}^{n}|a_{ij}|$,
4. *induced-norm :* $\|A\|_p = \sup_{x\in\mathbb{R}^n, x\neq 0}\frac{\|Ax\|_p}{\|x\|_p}$.

Corollary 1.7. *For all $A \in \mathbb{C}^{n\times n}$,*

$$\|A\|_2 \leq \|A\|_F \leq \sqrt{n}\|A\|_2, \tag{107}$$

$$\frac{1}{\sqrt{n}}\|A\|_2 \leq \|A\|_\infty \leq \sqrt{n}\|A\|_2, \tag{108}$$

$$\frac{1}{\sqrt{n}}\|A\|_\infty \leq \|A\|_2 \leq \sqrt{n}\|A\|_\infty, \tag{109}$$

$$\frac{1}{\sqrt{n}}\|A\|_1 \leq \|A\|_2 \leq \sqrt{n}\|A\|_1. \tag{110}$$

Corollary 1.8. *For all $A \in \mathbb{C}^{n\times n}$, then $\|A\|_2 \leq \sqrt{\|A\|_1\|A\|_\infty}$.*

Proof.

$$\|A\|_2^2 = \rho(A)^2 = \lambda \leq \|A\|_1\|A^*\|_1 = \|A\|_1\|A\|_\infty.$$

where λ is the eigenvalue of A^*A. □

Theorem 1.26. *(Matrix 2-norm and Frobenius invariance) (Matrix 2-norm and Frobenius are invariant under the orthogonal transformation, i.e., if Q is an $n\times n$ orthogonal matrix, then*

$$\|QA\|_2 = \|A\|_2, \ \forall A \in \mathbb{R}^{n\times n}, \tag{111}$$

$$\|QA\|_F = \|A\|_F, \ \forall A \in \mathbb{R}^{n\times n} \tag{112}$$

Theorem 1.27. *(Neumann Series) Suppose that $A \in \mathbb{R}^{n\times n}$. If $\|A\| < 1$, then $(I-A)$ is nonsingular and*

$$(I-A)^{-1} = \sum_{k=0}^{\infty} A^k \tag{113}$$

with

$$\frac{1}{1+\|A\|} \le \left\|(I-A)^{-1}\right\| \le \frac{1}{1-\|A\|}. \tag{114}$$

Moreover, if A is nonnegative, then $(I-A)^{-1} = \sum_{k=0}^{\infty} A^k$ is also nonnegative.

Proof. 1. (I-A) is nonsingular, i.e. $(I-A)^{-1}$ exits.

$$\begin{aligned} \left\|(I-A)x\right\| &\ge \|Ix\| - \|Ax\| \\ &\ge \|x\| - \|A\|\|x\| \\ &= (1-\|A\|)\|x\| \\ &= C\|x\|. \end{aligned}$$

So, we get if $(I-A)x = 0$, then $x = 0$. Therefore, $ker(I-A) = 0$, then $(I-A)^{-1}$ exists.

2. Let $S_N = \sum_{k=0}^{N} A^k$, we want to show $(I-A)S_N \to I$, as $N \to \infty$. First, we would like to show $\left\|A^k\right\| \le \|A\|^k$.

$$\left\|A^k\right\| = \sup_{0\neq x\in\mathbb{C}^n} \frac{\left\|A^k x\right\|}{\|x\|} \le \sup_{0\neq x\in\mathbb{C}^n} \frac{\|A\|\left\|A^{k-1}x\right\|}{\|x\|} \le \cdots \le \|A\|^k.$$

$$(I-A)S_N = S_N - AS_N = \sum_{k=0}^{N} A^k - \sum_{k=1}^{N+1} A^k = A^0 - A^{N+1} = I - A^{N+1}.$$

So

$$\left\|(I-A)S_N - I\right\| = \left\|-A^{N+1}\right\| \le \|A\|^{N+1}.$$

Since $\|A\| < 1$, then $\|A\|^{N+1} \to 0$. Therefore,

$$(I-A)\sum_{k=0}^{\infty} A^k = I.$$

and

$$(I-A)^{-1} = \sum_{k=0}^{\infty} A^k$$

3. bounded norm
Since

$$1 = \|I\| = \left\|(I-A) * (I-A)^{-1}\right\|.$$

So,

$$(1-\|A\|)\left\|(I-A)^{-1}\right\| \le 1 \le (1+\|A\|)\left\|(I-A)^{-1}\right\|.$$

Therefore,

$$\frac{1}{1+\|A\|} \leq \left\|(I-A)^{-1}\right\| \leq \frac{1}{1-\|A\|}.$$

□

Lemma 1.1. *Suppose that $A \in \mathbb{R}^{n\times n}$. If $(I-A)$ is singular, then $\|A\| \geq 1$.*

Proof. Converse-negative proposition of If $\|A\| < 1$, then $(I-A)$ is nonsingular. □

Theorem 1.28. *Let A be a nonnegative matrix. then $\rho(A) < 1$ if only if $I-A$ is nonsingular and $(I-A)^{-1}$ is nonnegative.*

Proof. 1. By theorem (1.27).

2. $\Leftarrow$ since $I-A$ is nonsingular and $(I-A)^{-1}$ is nonnegative, by the Perron- Frobenius theorem, there is a nonnegative eigenvector u associated with $\rho(A)$, which is an eigenvalue, i.e.

$$Au = \rho(A)u$$

or

$$(I-A)^{-1}u = \frac{1}{1-\rho(A)}u.$$

since $I-A$ is nonsingular and $(I-A)^{-1}$ is nonnegative, this show that $1-\rho(A) > 0$, which implies

$$\rho(A) < 1.$$

□

1.5 Problems

Problem 1.2. *(Prelim Jan. 2011#2) Let $A \in \mathbb{C}^{m\times n}$ and $b \in \mathbb{C}^m$. Prove that the vector $x \in \mathbb{C}^n$ is a least squares solution of $Ax = b$ if and only if $r \perp \text{range}(A)$, where $r = b - Ax$.*

Solution. We already know, $x \in \mathbb{C}^n$ is a least squares solution of $Ax = b$ if and only if

$$A^*Ax = A^*b.$$

and

$$\begin{aligned}(r, Ax) = (Ax) * r &= x^*A^*(b-Ax) \\ &= x^*(A^*b - A^*Ax)) \\ &= 0.\end{aligned}$$

Therefore, $r \perp \text{range}(A)$. The above way is invertible, hence we prove the result. ◂

Problem 1.3. *(Prelim Jan. 2011#3) Suppose $A, B \in \mathbb{R}^{n\times n}$ and A is non-singular and B is singular. Prove that*

$$\frac{1}{\kappa(A)} \leq \frac{\|A-B\|}{\|A\|},$$

where $\kappa(A) = \|A\| \cdot \left\|A^{-1}\right\|$, and $\|\cdot\|$ is an reduced matrix norm.

Solution. Since B is singular, then there exists a vector $x \neq 0$, s.t. $Bx = 0$. Since A is non-singular, then A^{-1} is also non-singular. Moreover, $A^{-1}Bx = 0$. Then, we have

$$x = x - A^{-1}Bx = (I - A^{-1}B)x.$$

So

$$\|x\| = \left\|(I - A^{-1}B)x\right\| \leq \left\|A^{-1}A - A^{-1}B\right\| \|x\| \leq \left\|A^{-1}\right\| \|A-B\| \|x\|.$$

Since $x \neq 0$, so

$$1 \leq \left\|A^{-1}\right\| \|A-B\|.$$

$$\frac{1}{\left\|A^{-1}\right\| \|A\|} \leq \frac{\|A-B\|}{\|A\|},$$

i.e.

$$\frac{1}{\kappa(A)} \leq \frac{\|A-B\|}{\|A\|}.$$

◀

Problem 1.4. *(Prelim Aug. 2010#2) Suppose that $A \in \mathbb{R}^{n\times n}$ is SPD.*

1. *Show that $\|x\|_A = \sqrt{x^T Ax}$ defines a vector norm.*
2. *Let the eigenvalues of A be ordered so that $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. Show that*

$$\sqrt{\lambda_1}\|x\|_2 \leq \|x\|_A \leq \sqrt{\lambda_n}\|x\|_2.$$

   for any $x \in \mathbb{R}^n$.

3. *Let $b \in \mathbb{R}^n$ be given. Prove that $x_* \in \mathbb{R}^n$ solves $Ax = b$ if and only if x_* minimizes the quadratic function $f : \mathbb{R}^n \to \mathbb{R}$ defined by*

$$f(x) = \frac{1}{2}x^T Ax - x^T b.$$

Solution. 1. (a) Obviously, $\|x\|_A = \sqrt{x^T Ax} \geq 0$. When $x = 0$, then $\|x\|_A = \sqrt{x^T Ax} = 0$; when $\|x\|_A = \sqrt{x^T Ax} = 0$, then we have $(Ax, x) = 0$, since A is SPD, therefore, $x \equiv 0$.

(b) $\|\lambda x\|_A = \sqrt{\lambda x^T A \lambda x} = \sqrt{\lambda^2 x^T Ax} = |\lambda|\sqrt{x^T Ax} = |\lambda|\|x\|_A$.

(c) Next we will show $\left\|x+y\right\|_A \leq \|x\|_A + \left\|y\right\|_A$. First, we would like to show

$$\left|y^T Ax\right| \leq \|x\|_A \left\|y\right\|_A.$$

Since A is SPD, therefore $A = R^T R$, moreover

$$\|Rx\|_2 = (Rx, Rx)^{1/2} = \sqrt{(Rx)^T Rx} = \sqrt{x^T R^T Rx} = \sqrt{x^T Ax} = \|x\|_A .$$

Then

$$\left|y^T Ax\right| = \left|y^T R^T Rx\right| = \left|(Ry)^T Rx\right| = \left|(Rx, Ry)\right| \overset{c.s.}{\leq} \|Rx\|_2 \left\|Ry\right\|_2 = \|x\|_A \left\|y\right\|_A .$$

And

$$\begin{aligned} \left\|x+y\right\|_A^2 &= (x+y, x+y)_A = (x,x)_A + 2(x,y)_A + (y,y)_A \\ &\leq \|x\|_A + 2\left|y^T Ax\right| + \left\|y\right\|_A \\ &\leq \|x\|_A + 2\|x\|_A \left\|y\right\|_A + \left\|y\right\|_A \\ &= \left(\|x\|_A + \left\|y\right\|_A\right)^2 . \end{aligned}$$

therefore

$$\left\|x+y\right\|_A \leq \|x\|_A + \left\|y\right\|_A .$$

2. Since A is SPD, therefore $A = R^T R$, moreover

$$\|Rx\|_2 = (Rx, Rx)^{1/2} = \sqrt{(Rx)^T Rx} = \sqrt{x^T R^T Rx} = \sqrt{x^T Ax} = \|x\|_A .$$

Let $0 < \tilde{\lambda}_1 \leq \tilde{\lambda}_2 \leq \cdots \leq \tilde{\lambda}_n$ be the eigenvalue of R, then $\tilde{\lambda}_i = \sqrt{\lambda_i}$. so

$$\left|\tilde{\lambda}_1\right| \|x\|_2 \leq \|Rx\|_2 = \|x\|_A \leq \left|\tilde{\lambda}_n\right| \|x\|_2 .$$

i.e.

$$\sqrt{\lambda_1}\, \|x\|_2 \leq \|Rx\|_2 = \|x\|_A \leq \sqrt{\lambda_n}\, \|x\|_2 .$$

3. Since

$$\begin{aligned} \frac{\partial}{\partial x_i}\left(x^T Ax\right) &= \frac{\partial}{\partial x_i}\left(x^T\right) Ax + x^T \frac{\partial}{\partial x_i}(Ax) \\ &= [0, \cdots, 0, \underset{i}{1}, 0, \cdots, 0] Ax + x^T A \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} i \\ &= (Ax)_i + \left(A^T x\right)_i = 2(Ax)_i . \end{aligned}$$

and

$$\frac{\partial}{\partial x_i}\left(x^T b\right) = \frac{\partial}{\partial x_i}\left(x^T\right) b = [0, \cdots, 0, \underset{i}{1}, 0, \cdots, 0] b = b_i .$$

Therefore,

$$\nabla f(x) = \frac{1}{2} 2Ax - b = Ax - b.$$

If $Ax_* = b$, then $\nabla f(x_*) = Ax_* - b = 0$, therefore x_* minimizes the quadratic function f. Conversely, when x_* minimizes the quadratic function f, then $\nabla f(x_*) = Ax_* - b = 0$, therefore $Ax_* = b$.

◀

2 Direct Method

2.1 For squared or rectangular matrices $A \in \mathbb{C}^{m,n}, m \geq n$

2.1.1 Singular Value Decomposition

Theorem 2.1. *(Reduced SVD) Suppose that $A \in \mathbb{R}^{m\times n}$.*

$$A = \underbrace{\hat{U}}_{m\times n}\ \underbrace{\hat{\Sigma}}_{n\times n}\ \underbrace{\hat{V}^*}_{n\times n}.$$

This is called a Reduced SVD of A. where

- *σ_i– Singular values and $\hat{\Sigma} = diag(\sigma_1, \sigma_2, \cdots, \sigma_n) \in \mathbb{R}^{n\times n}$.*
- *v_i– right singular vectors and $\hat{U} = [u_1, u_2, \cdots, u_n]$.*
- *u_i– left singular vectors and $\hat{V} = [v_1, v_2, \cdots, v_n]$.*

Theorem 2.2. *(SVD) Suppose that $A \in \mathbb{R}^{m\times n}$.*

$$A = \underbrace{U}_{m\times m}\ \underbrace{\Sigma}_{m\times n}\ \underbrace{V^*}_{n\times n}.$$

This is called a SVD of A. where

- *σ_i– Singular values and $\hat{\Sigma} = diag(\sigma_1, \sigma_2, \cdots, \sigma_n) \in \mathbb{R}^{n\times n}$.*
- *v_i– right singular vectors, $\hat{U} = [u_1, u_2, \cdots, u_m]$ and U is unitary.*
- *u_i– left singular vectors, $\hat{V} = [v_1, v_2, \cdots, v_n]$ and V is unitary.*

Remark 2.1. *1. SVD works for any matrices, spectral decomposition only works for squared matrices.*

2. The spectral decomposition $A = X\Lambda X^{-1}$ works only if A is non-defective matrices.

For a symmetric matrix the following decompositions are equivalent to SVD.

1. *Eigen-value decomposition: i.e. $A = X\Sigma X^{-1}$. When A is symmetric, the eigenvalues are real and the eigenvectors can be chosen to be orthonormal and hence $X^TX = XX^T = I$ i.e.$X^{-1} = XT$. The only difference is that the singular values are the magnitudes of the eigenvalues and hence the column of X needs to be multiplied by a negative sign if the eigenvalue turns out to be negative to get the singular value decomposition. Hence, U=X and $\sigma_i = |\lambda_i|$.*
2. *Orthogonal decomposition: i.e. $A = PDP^T$, where P is a unitary matrix and D is a diagonal matrix. This exists only when matrix A is symmetric and is the same as eigenvalue decomposition.*
3. *Schur decomposition i.e. $A = QSQ^T$, where Q is a unitary matrix and S is an upper triangular matrix. This can be done for any matrix. When A is symmetric, then S is a diagonal matrix and again is the same as the eigenvalue decomposition and orthogonal decomposition.*

2.1.2 Gram-Schmidt orthogonalization

Definition 2.1. *(projection operator) We define the projection operator as*

$$\text{proj}_{\mathbf{u}}(\mathbf{v}) = \frac{(\mathbf{u},\mathbf{v})}{(\mathbf{u},\mathbf{u})}\mathbf{u},$$

where $(\mathbf{u},\mathbf{v})$ *is the inner product of the vector* $\mathbf{u}$ *and* $\mathbf{v}$*. If* $\mathbf{u} = \mathbf{0}$*, we define*

$$\text{proj}_0(\mathbf{v}) = \mathbf{0}.$$

Remark 2.2. *1. This operator projects the vector* $\mathbf{v}$ *orthogonally onto the line spanned by vector* $\mathbf{u}$*.*

2. the projection map proj_0 *is the zero map, sending every vector to the zero vector.*

Definition 2.2. *(Gram-Schmidt orthogonalization) The Gram-Schmidt process then works as follows:*

$$\begin{aligned}
\mathbf{u}_1 &= \mathbf{v}_1, & \mathbf{q}_1 &= \frac{\mathbf{u}_1}{\|\mathbf{u}_1\|}\\
\mathbf{u}_2 &= \mathbf{v}_2 - \text{proj}_{\mathbf{u}_1}(\mathbf{v}_2), & \mathbf{q}_2 &= \frac{\mathbf{u}_2}{\|\mathbf{u}_2\|}\\
\mathbf{u}_3 &= \mathbf{v}_3 - \text{proj}_{\mathbf{u}_1}(\mathbf{v}_3) - \text{proj}_{\mathbf{u}_2}(\mathbf{v}_3), & \mathbf{q}_3 &= \frac{\mathbf{u}_3}{\|\mathbf{u}_3\|}\\
\mathbf{u}_4 &= \mathbf{v}_4 - \text{proj}_{\mathbf{u}_1}(\mathbf{v}_4) - \text{proj}_{\mathbf{u}_2}(\mathbf{v}_4) - \text{proj}_{\mathbf{u}_3}(\mathbf{v}_4), & \mathbf{q}_4 &= \frac{\mathbf{u}_4}{\|\mathbf{u}_4\|}\\
&\vdots & &\vdots\\
\mathbf{u}_k &= \mathbf{v}_k - \sum_{j=1}^{k-1}\text{proj}_{\mathbf{u}_j}(\mathbf{v}_k), & \mathbf{q}_k &= \frac{\mathbf{u}_k}{\|\mathbf{u}_k\|}.
\end{aligned}$$

$$A = [a_1, a_2, \cdots, a_n] = [q_1, q_2, \cdots, q_n]\begin{bmatrix} r_{11} & \cdots & r_{1n} \\ & r_{22} & \vdots \\ & & r_{nn} \end{bmatrix}.$$

Definition 2.3. *(projector) A projector is a square matrix* P *that satisfies*

$$P^2 = P.$$

Definition 2.4. *(complementary projector) If* P *is a projector, then*

$$I - P$$

is also a projector and is called complementary projector.

Definition 2.5. *(orthogonal projector) If* P *is a orthogonal projector if only if*

$$P = P^*.$$

The complement of an orthogonal projector is also orthogonal projector.

Definition 2.6. *(projection with orthonormal basis) If P is a orthogonal projector, then $P = P^*$ and P has SVD, i.e. $P = Q\Sigma Q^*$. Since an orthogonal projector has some singular values equal to zero (except the identity map P=I), it is natural to drop the silent columns of Q and use the reduced rather than full SVD, i.e.*

$$P = \hat{Q}\hat{Q}^*.$$

The complement projects onto the space orthogonal to range($\hat{Q}$).

Definition 2.7. *(Gram- Schmidt projections)*

$$P = I - \hat{Q}\hat{Q}^*.$$

The complement projects onto the space orthogonal to range($\hat{Q}$).

Definition 2.8. *(Householder reflectors) The householder reflector F is a particular matrix which satisfies*

$$F = I - 2\frac{vv^*}{\|v\|}.$$

Comparsion 2.1. *(Gram- Schmidt and Householder)*

$Gram-Schmidt$	$\underbrace{AR_1R_2\cdots R_n}_{\hat{R}^{-1}} = \hat{Q}$	*triangular orthogonalization*
$Householder$	$\underbrace{Q_n\cdots Q_2Q_1}_{Q^*}A = R$	*orthogonal triangularization*

2.1.3 QR Decomposition

Theorem 2.3. *(Reduced QR Decomposition) Suppose that $A \in \mathbb{C}^{m\times n}$.*

$$A = \underset{m\times n}{\hat{Q}}\ \underset{n\times n}{\hat{R}}.$$

This is called a Reduced QR Decomposition of A. where

- $\hat{Q} \in \mathbb{C}^{m\times n}$ *– with orthonormal columns.*
- $\hat{R} \in \mathbb{C}^{n\times n}$ *– upper triangular matrix.*

Theorem 2.4. *(QR Decomposition) Suppose that $A \in \mathbb{C}^{m\times n}$.*

$$A = \underset{m\times m}{Q}\ \underset{m\times n}{R}.$$

This is called a QR Decomposition of A. where

- $Q \in \mathbb{C}^{m\times m}$ *– is unitary.*
- $R \in \mathbb{C}^{m\times n}$ *– upper triangular matrix.*

Theorem 2.5. *(Existence of QR Decomposition) Every $A \in \mathbb{C}^{m\times n}$ has full and reduced QR decomposition.*

Theorem 2.6. *(Uniqueness of QR Decomposition) Each $A \in \mathbb{C}^{m\times n}$ of full rank has a unique reduced QR decomposition $A = \hat{Q}\hat{R}$ with $r_{jj} > 0$.*

2.2 For squared matrices $A \in \mathbb{C}^{n,n}$

A problem can be read as

$$\begin{array}{lccc} f: & D & \to & S \\ & Data & \to & Solution \end{array}$$

2.2.1 Condition number

Definition 2.9. *(Well posedness) We say that a problem is well- posed if the solution depends continuously on the data, otherwise we say it is ill-posed.*

Definition 2.10. *(absolute condition number) The absolute condition number $\hat{\kappa} = \hat{\kappa}(x)$ of the problem f at x is defined as*

$$\hat{\kappa} = \lim_{\delta\to 0} \sup_{\|\delta x\|\le\delta} \frac{\|f(x+\delta x) - f(x)\|}{\|\delta x\|}.$$

If f is (Freēchet) differentiable

$$\hat{\kappa} = \|Df(x)\|.$$

Example 2.1. *f: $R^2 \to R$ and $f(x_1, x_2) = x_1 - x_2$, then*

$$Df(x) = [\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}] = [1, -1].$$

and

$$\hat{\kappa} = \|Df(x)\|_\infty = 1.$$

Definition 2.11. *(relative condition number) The absolute condition number $\kappa = \kappa(x)$ of the problem f at x is defined as*

$$\begin{aligned} \kappa &= \lim_{\delta\to 0} \sup_{\|\delta x\|\le\delta} \frac{\frac{\|f(x+\delta x)-f(x)\|}{\|\delta x\|}}{\frac{\|f(x)\|}{\|x\|}} \\ &= \frac{\|x\|}{\|f(x)\|}\hat{\kappa}. \end{aligned}$$

Definition 2.12. *(condition number of Matrix- Vector Multiplication) The absolute condition of $f(x) = Ax$*

$$\begin{aligned}\kappa &= \lim_{\delta\to 0}\sup_{\|\delta x\|\le\delta}\frac{\frac{\|A(x+\delta x)-A(x)\|}{\|\delta x\|}}{\frac{\|A(x)\|}{\|x\|}}\\ &= \|A\|\frac{\|x\|}{\|A(x)\|}.\end{aligned}$$

Theorem 2.7. *(condition of Matrix- Vector Multiplication) Since, $\|x\| = \|A^{-1}Ax\| \le \|A^{-1}\|\|Ax\|$, then $\frac{\|x\|}{\|A(x)\|} \le \|A^{-1}\|$. So*

$$\kappa \le \|A\|\|A^{-1}\|.$$

Particularly,

$$\kappa = \|A\|_2\|A^{-1}\|_2.$$

Definition 2.13. *(condition number of Matrix) Let $A \in \mathbb{C}^{n\times n}$, invertible, the condition number of A is*

$$\kappa(A)_{\|\cdot\|} = \|A\|\|A^{-1}\|.$$

particularly,

$$\kappa_2(A) = \|A\|_2\|A^{-1}\|_2 = \frac{\sigma_1}{\sigma_n}.$$

where $\sigma_1\cdots\sigma_n$ are singular value of A. So, $\|A\|_2 = \sigma_1$.

2.2.2 LU Decomposition

Definition 2.14. *(LU Decomposition without pivoting) Let $A \in \mathbb{C}^{n\times n}$. An LU factorization refers to the factorization of A, with proper row and/or column orderings or permutations, into two factors, a lower triangular matrix L and an upper triangular matrix U ,*

$$A = LU.$$

In the lower triangular matrix all elements above the diagonal are zero, in the upper triangular matrix, all the elements below the diagonal are zero. For example, for a 3-by-3 matrix A, its LU decomposition looks like this:

$$\begin{bmatrix}a_{11} & a_{12} & a_{13}\\ a_{21} & a_{22} & a_{23}\\ a_{31} & a_{32} & a_{33}\end{bmatrix} = \begin{bmatrix}l_{11} & 0 & 0\\ l_{21} & l_{22} & 0\\ l_{31} & l_{32} & l_{33}\end{bmatrix}\begin{bmatrix}u_{11} & u_{12} & u_{13}\\ 0 & u_{22} & u_{23}\\ 0 & 0 & u_{33}\end{bmatrix}.$$

Definition 2.15. *(LU Decomposition with partial pivoting) The LU factorization with Partial Pivoting refers often to the LU factorization with row permutations only,*

$$PA = LU,$$

where L and U are again lower and upper triangular matrices, and P is a permutation matrix which, when left-multiplied to A, reorders the rows of A.

Definition 2.16. *(LU Decomposition with full pivoting) An LU factorization with full pivoting involves both row and column permutations,*

$$PAQ = LU,$$

where L, U and P are defined as before, and Q is a permutation matrix that reorders the columns of A

Definition 2.17. *(LDU Decomposition) An LDU decomposition is a decomposition of the form*

$$A = \tilde{L}D\tilde{U},$$

where D is a diagonal matrix and L and U are unit triangular matrices , meaning that all the entries on the diagonals of L and U are one.

For example, for a 3-by-3 matrix A, its LDU decomposition looks like this:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{bmatrix} \begin{bmatrix} 1 & u_{12} & u_{13} \\ 0 & 1 & u_{23} \\ 0 & 0 & 1 \end{bmatrix}.$$

Theorem 2.8. *(existence of Decomposition) Any square matrix A admits an LUP factorization. If A is invertible, then it admits an LU (or LDU) factorization if and only if all its leading principal minors are nonsingular. If A is a singular matrix of rank k , then it admits an LU factorization if the first k leading principal minors are nonsingular, although the converse is not true.*

2.2.3 Cholesky Decomposition

Definition 2.18. *(Cholesky Decomposition) In linear algebra, the Cholesky decomposition or Cholesky factorization is a decomposition of a Hermitian, positive-definite matrix into the product of a lower triangular matrix and its conjugate transpose,*

$$A = LL^*.$$

Definition 2.19. *(LDM Decomposition) Let $A \in \mathbb{R}^{n\times n}$ and all the leading principal minors $det(A(1:k;1:k)) \neq 0; k = 1,\cdots,n-1$. Then there exist unique unit lower triangular matrices L and M and a unique diagonal matrix $D = diag(d_1,\cdots,d_n)$, such that*

$$A = \tilde{L}DM^T.$$

Definition 2.20. *(LDL Decomposition) A closely related variant of the classical Cholesky decomposition is the LDL decomposition,*

$$A = \tilde{L}D\tilde{L}^*,$$

where L is a lower unit triangular (unitriangular) matrix and D is a diagonal matrix.

Remark 2.3. *This decomposition is related to the classical Cholesky decomposition, of the form LL^*, as follows:*

$$A = \tilde{L}D\tilde{L}^* = \tilde{L}D^{\frac{1}{2}}D^{\frac{1}{2}*}\tilde{L}^* = \tilde{L}D^{\frac{1}{2}}(\tilde{L}D^{\frac{1}{2}})^*.$$

The LDL variant, if efficiently implemented, requires the same space and computational complexity to construct and use but avoids extracting square roots. Some indefinite matrices for which no Cholesky decomposition exists have an LDL decomposition with negative entries in D. For these reasons, the LDL decomposition may be preferred. For real matrices, the factorization has the form $A = LDL^T$ and is often referred to as LDL^T decomposition (or LDL^T decomposition). It is closely related to the eigendecomposition of real symmetric matrices, $A = Q\Lambda Q^T$.

2.2.4 The Relationship of the Existing Decomposition

From last subsection, If $A = A^*$, then

1. diagonal elements of A are real and positive .
2. principal sub matrices of A are HPD .

Comparsion 2.2. *(Gram- Schmidt and Householder)*

$A = \tilde{L}DM^*$	$A = \tilde{L}DM^* = \tilde{L}D\tilde{L}^*$	$\tilde{L} = M$
$A = \tilde{L}D\tilde{L}^*$	$A = \tilde{L}D\tilde{L}^* = \tilde{L}D^{\frac{1}{2}}D^{\frac{1}{2}*}\tilde{L}^*$	$L = \tilde{L}D^{\frac{1}{2}}$
$A = LU$	$A = LU = LL^*$	$U = L^*$

2.2.5 Regular Splittings[3]

Definition 2.21. *(Regular Splittings) Let A, M, N be three given matrices satisfying*

$$A = M - N.$$

The pair of matrices M, N is a regular splitting of A, if M is nonsingular and M^{-1} and N are nonnegative .

Theorem 2.9. *(The eigenvalue radius estimation of Regular Splittings[3]) Let M, N be a regular splitting of A. Then*

$$\rho(M^{-1}N) < 1$$

if only if A is nonsingular and A^{-1} is nonnegative.

Proof. 1. Define $G = M^{-1}N$, since $\rho(G) < 1$, then $I - G$ is nonsingular. And then $A = M(I - G)$, so A is nonsingular. So, by Theorem.1.28 satisfied, since $G = M^{-1}N$ is nonsingular and $\rho(G) < 1$, then we have $(I - G)^{-1}$ is nonnegative as is $A^{-1} = (I - G)^{-1}M^{-1}$.

2. $\Leftarrow$: since A,M are nonsingular and A^{-1} is nonnegative, then $A = M(I-G)$ is nonsingular. Moreover

$$\begin{aligned} A^{-1}N &= \left(M(I-M^{-1}N)\right)^{-1}N \\ &= (I-M^{-1}N)^{-1}M^{-1}N \\ &= (I-G)^{-1}G. \end{aligned}$$

Clearly, $G = M^{-1}N$ is nonnegative by the assumptions, and as a result of the Perron-Frobenius theorem, there is a nonnegative eigenvector x associated with $\rho(G)$ which is an eigenvalue, such that

$$Gx = \rho(G)x.$$

Therefore

$$A^{-1}Nx = \frac{\rho(G)}{1-\rho(G)}x.$$

Since x and $A^{-1}N$ are nonnegative, this shows that

$$\frac{\rho(G)}{1-\rho(G)} \geq 0.$$

and this can be true only when $0 \leq \rho(G) \leq 1$. Since $I-G$ is nonsingular, then $\rho(G) \neq 1$, which implies that $\rho(G) < 1$.

□

2.3 Problems

Problem 2.1. *(Prelim Aug. 2010#1) Prove that $A \in \mathbb{C}^{m\times n}(m > n)$ and let $A = \hat{Q}\hat{R}$ be a reduced QR factorization.*

1. *Prove that A has rank n if and only if all the diagonal entries of $\hat{R}$ are non-zero.*
2. *Suppose* $\text{rank}(A) = n$, *and define* $P = \hat{Q}\hat{Q}^*$. *Prove that* $\text{range}(P) = \text{range}(A)$.
3. *What type of matrix is P?*

Solution. 1. From the properties of reduced QR factorization, we know that $\hat{Q}$ has orthonormal columns, therefore $\det(\hat{Q}) = 1$ and $\hat{R}$ is upper triangular matrix, so $\det(\hat{R}) = \prod_{i=1}^{n} r_{ii}$. Then

$$\det(A) = \det(\hat{Q}\hat{R}) = \det(\hat{Q})\det(\hat{R}) = \prod_{i=1}^{n} r_{ii}.$$

Therefore, A has rank n if and only if all the diagonal entries of $\hat{R}$ are non-zero.

2. (a) $\text{range}(A) \subseteq \text{range}(P)$: Let $y \in \text{range}(A)$, that is to say there exists a $x \in \mathbb{C}^n$ s.t. $Ax = y$. Then by reduced QR factorization we have $y = \hat{Q}\hat{R}x$. then

$$Py = P\hat{Q}\hat{R}x = \hat{Q}\hat{Q}^*\hat{Q}\hat{R}x = \hat{Q}\hat{R}x = Ax = y.$$

therefore $y \in \text{range}(P)$.

(b) $\text{range}(P) \subseteq \text{range}(A)$: Let $v \in \text{range}(P)$, that is to say there exists a $v \in \mathbb{C}^n$, s.t. $v = Pv = \hat{Q}\hat{Q}^*v$.

Claim 2.1.

$$\hat{Q}\hat{Q}^* = A\left(A^*A\right)^{-1}A^*.$$

Proof.

$$\begin{aligned} A\left(A^*A\right)^{-1}A^* &= \hat{Q}\hat{R}\left(\hat{R}^*\hat{Q}^*\hat{Q}\hat{R}\right)^{-1}\hat{R}^*\hat{Q}^* \\ &= \hat{Q}\hat{R}\left(\hat{R}^*\hat{R}\right)^{-1}\hat{R}^*\hat{Q}^* \\ &= \hat{Q}\hat{R}\hat{R}^{-1}\left(\hat{R}^*\right)^{-1}\hat{R}^*\hat{Q}^* \\ &= \hat{Q}\hat{Q}^*. \end{aligned}$$

◀

Therefore by the claim, we have

$$v = Pv = \hat{Q}\hat{Q}^*v = A\left(A^*A\right)^{-1}A^*v = A\left(\left(A^*A\right)^{-1}A^*v\right) = Ax.$$

where $x = \left(A^*A\right)^{-1}A^*v$. Hence $v \in \text{range}(A)$.

3. P is an orthogonal projector.

◀

Problem 2.2. *(Prelim Aug. 2010#4) Prove that $A \in \mathbb{R}^{n\times n}$ is SPD if and only if it has a Cholesky factorization.*

Solution. 1. Since A is SPD, so it has LU factorization, and $L = U$, i.e.

$$A = LU = U^TU.$$

Therefore, it has a Cholesky factorization.

2. if A has Cholesky factorization, i.e $A = U^TU$, then

$$x^TAx = x^TU^TUx = (Ux)^TUx.$$

Let $y = Ux$, then we have

$$x^TAx = (Ux)^TUx = y^Ty = y_1^2 + y_2^2 + \cdots + y_n^2 \geq 0,$$

with equality only when y = 0, i.e. x=0 (since U is non-singular). Hence A is SPD.

◀

Problem 2.3. *(Prelim Aug. 2009#2) Prove that for any matrix $A \in \mathbb{C}^{n\times n}$, singular or nonsingular, there exists a permutation matrix $P \in \mathbb{R}^{n\times n}$ such that PA has an LU factorization, i.e. PA=LU.*

Solution. ◀

Problem 2.4. *(Prelim Aug. 2009#4) Let $A \in \mathbb{C}^{n\times n}$ and $\sigma_1 \geq \sigma_2 \geq \cdots \sigma_n \geq 0$ be its singular values.*

1. *Let λ be an eigenvalue of A. Show that $|\lambda| \leq \sigma_1$.*
2. *Show that $\left|\det(A)\right| = \prod_{j=1}^{n}\sigma_j$.*

Solution. 1. Since $\sigma_1 = \|A\|_2$(proof follows by induction), so we need to show $|\lambda| \leq \|A\|_2$.

$$|\lambda|\|x\|_2 = \|\lambda x\|_2 = \|Ax\| \leq \|A\|_2\|x\|_2.$$

Therefore,

$$|\lambda| \leq \sigma_1.$$

2.

$$\left|\det(A)\right| = \left|\det(U\Sigma V^*)\right|\left|\det(U)\right|\left|\det(\Sigma)\right|\left|\det(V^*)\right| = \left|\det(\Sigma)\right| = \prod_{j=1}^{n}\sigma_j.$$

◀

Problem 2.5. *(Prelim Aug. 2009#4) Let*

Solution. ◀

3 Iterative Method

3.1 Diagonal dominant

Definition 3.1. *(Diagonal dominant of size δ)* *$A \in \mathbb{C}^{n\times n}$ has diagonal dominant of size $\delta > 0$ if*

$$|a_{ii}| \geq \sum_{j\neq i} |a_{ij}| + \delta.$$

Properties 3.1. *If $A \in \mathbb{C}^{n\times n}$ is diagonal dominant of size $\delta > 0$ then*

1. *A^{-1} exists.*
2. *$\left\|A^{-1}\right\|_\infty \leq \frac{1}{\delta}$.*

Proof. 1. Let $b = Ax$ and chose $k \in (1,2,3,\cdots,n)$ s.t $\|x\|_\infty = |x_k|$. Moreover, let $b_k = \sum_{j=1}^n a_{kj}x_j$. Since

$$|a_{ii}| \geq \sum_{j\neq i} |a_{ij}| + \delta,$$

and

$$\sum_{j\neq i} |a_{ij}x_j| \leq \sum_{j\neq i} |a_{ij}||x_j| \leq \|x\|_\infty \sum_{j\neq i} |a_{ij}|.$$

Then

$$\begin{aligned}
|b_k| &= \left|\sum_{j=1}^n a_{kj}x_j\right| \\
&= \left|a_{kk}x_k + \sum_{j\neq k}^n a_{kj}x_j\right| \\
&\geq |a_{kk}x_k| - \left|\sum_{j\neq k}^n a_{kj}x_j\right| \\
&\geq |a_{kk}x_k| - \|x\|_\infty \sum_{j\neq i} |a_{ij}| \\
&\geq |a_{kk}|\|x\|_\infty - \|x\|_\infty \sum_{j\neq i} |a_{ij}| \\
&= \delta\|x\|_\infty.
\end{aligned}$$

So, $\|Ax\|_\infty = \|b\|_\infty \geq \|b_k\|_\infty \geq \delta\|x\|_\infty$. If $Ax = 0$, then $x = 0$. So, $ker(A) = 0$, and then, A^{-1} exists.

2. Since $\|Ax\|_\infty = \|b\|_\infty \geq \|b_k\|_\infty \geq \delta\|x\|_\infty$, so $\|Ax\| \geq \delta\|x\|_\infty$ and $\left\|A^{-1}\right\|_\infty \leq \frac{1}{\delta}$.

□

3.2 General Iterative Scheme

An iterative scheme for the solution

$$Ax = b, \tag{115}$$

is a sequence given by

$$x^{k+1} = \phi(A, b, x^k, \cdots, x^{k-r}).$$

1. $r = 0$- two layer scheme.
2. $r \geq 1$multi-layer scheme.
3. ϕ - is a linear function of its arguments then the scheme is linear, otherwise it is nonlinear.
4. convergent if $x_k \overset{k\to\infty}{\to} x$.

Definition 3.2. *(General Iterative Scheme) A general linear two layer iterative scheme reads*

$$B_k\left(\frac{x^{k+1} - x^k}{\alpha_k}\right) + Ax^k = b.$$

1. *$\alpha_k \in R, B_k \in \mathbb{C}^{n\times n}$–iterative parameters*
2. *If $\alpha_k = \alpha, B_k = B$, then the method is stationary.*
3. *If $B_k = I$, then the method is explicit.*

If $x^k \to x_0$, then x_0 solves $Ax = b$. So

$$B_k\left(\frac{x_0 - x_0}{\alpha_k}\right) + Ax_0 = b,$$

i.e.

$$Ax_0 = b.$$

Now, consider the stationary scheme, i.e

$$B\left(\frac{x^{k+1} - x^k}{\alpha}\right) + Ax^k = b.$$

Then we get

$$x^{k+1} = x^k + \alpha B^{-1}(b - Ax^k).$$

Definition 3.3. *(Error Transfer Operator) Let $e^k = x - x^k$, where x is exact solution and x^k is the approximate solution at k step. Then*

$$\begin{aligned} x &= x + \alpha B^{-1}(b - Ax) \\ x^{k+1} &= x^k + \alpha B^{-1}(b - Ax^k). \end{aligned}$$

So, we get

$$e^{k+1} = e^k + \alpha B^{-1} Ae^k = (I - \alpha B^{-1} A)e^k := Te^k.$$

$T = I - \alpha B^{-1} A$ is the error transfer operator.

After we defined the error transfer operator , the iterative can be written as

$$x^{k+1} = Tx^k + \alpha B^{-1} b.$$

Theorem 3.1. *(sufficient condition for converges) The sufficient condition for converges is*

$$\|T\| < 1. \tag{116}$$

Theorem 3.2. *(sufficient & necessary condition for converges) The sufficient & necessary condition for converges is*

$$\rho(T) < 1, \tag{117}$$

where $\rho(T)$ is the spectral radius of T.

3.3 Stationary cases iterative method

3.3.1 Jacobi Method

Definition 3.4. *(Jacobi Method) Let*

$$A = L + D + U.$$

A Jacobi Method scheme reads

$$D\left(x^{k+1} - x^k\right) + Ax^k = b.$$

i.e. $\alpha_k = 1, B = D$ in the general iterative scheme.

Definition 3.5. *(Error Transfer Operator for Jacobi Method) the error transfer operator for Jacobi Method is as follows*

$$T = I - D^{-1}A.$$

Remark 3.1. *Since*

$$A = L + D + U.$$

and

$$D\left(x^{k+1} - x^k\right) + Ax^k = b.$$

Then we have

$$D\left(x^{k+1} - x^k\right) + (L + D + U)x^k = Lx^k + Dx^{k+1} + Ux^k = b.$$

So, the Jacobi iterative method can be written as

$$\sum_{j<i} a_{ij}x_j^k + a_{ii}x_i^{k+1} + \sum_{j>i} a_{ij}x_j^k = b_i,$$

or

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

Theorem 3.3. *(convergence of the Jacobi Method) If A is diagonal dominant , then the Jacobi Method convergences.*

Proof. We want to show If A is diagonal dominant , then $\|T_J\| < 1$, then Jacobi Method convergences. From the definition of T, we know that T for Jacobi Method is as follows

$$T_J = I - D^{-1}A.$$

In the matrix form is

$$T = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix} - \begin{pmatrix} \frac{1}{a_{11}} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{a_{nn}} \end{pmatrix} \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} = \left[t_{ij}\right] = \begin{cases} t_{ij} = 0, \ i = j, \\ t_{ij} = -\frac{a_{ij}}{a_{ii}}, \ i \neq j. \end{cases}.$$

So,

$$\|T\|_\infty = \max_i \sum_j |t_{ij}| = \max_i \sum_{i \neq j} |\frac{a_{ij}}{a_{ii}}|.$$

Since A is diagonal dominant, so

$$|a_{ii}| \geq \sum_{j \neq i} |a_{ij}| + \delta.$$

Therefore,

$$1 \geq \sum_{j \neq i} \frac{|a_{ij}|}{|a_{ii}|} + \frac{\delta}{|a_{ii}|}.$$

Hence, $\|T\|_\infty < 1$ □

3.3.2 Gauss-Seidel Method

Definition 3.6. *(Gauss-Seidel Method) Let*

$$A = L + D + U.$$

A Gauss-Seidel Method scheme reads

$$(L + D)\left(x^{k+1} - x^k\right) + Ax^k = b.$$

i.e. $\alpha_k = 1, B = L + D$ in the general iterative scheme.

Definition 3.7. *(Error Transfer Operator for Gauss-Seidel Method) The error transfer operator for Gauss-Seidel Method is as follows*

$$\begin{aligned} T &= I - (L + D)^{-1}A \\ &= I - (L + D)^{-1}(L + D + U) \\ &= -(L + D)^{-1}U. \end{aligned}$$

Remark 3.2. *The Gauss-Seidel method is an iterative technique for solving a square system of n linear equations with unknown x:*

$$A\mathbf{x} = \mathbf{b}.$$

It is defined by the iteration

$$L_*\mathbf{x}^{(k+1)} = \mathbf{b} - U\mathbf{x}^{(k)},$$

where the matrix A is decomposed into alower triangular component L_*, and a strictly upper triangular component U: $A = L_* + U$.

In more detail, write out A, x and b in their components:

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}, \qquad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}.$$

Then the decomposition of A into its lower triangular component and its strictly upper triangular component is given by:

$$A = L_* + U \qquad \text{where} \qquad L_* = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ a_{21} & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}, \quad U = \begin{bmatrix} 0 & a_{12} & \cdots & a_{1n} \\ 0 & 0 & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}.$$

The system of linear equations may be rewritten as:

$$L_*\mathbf{x} = \mathbf{b} - U\mathbf{x}$$

The Gauss-Seidel method now solves the left hand side of this expression for x, using previous value for x on the right hand side. Analytically, this may be written as:

$$\mathbf{x}^{(k+1)} = L_*^{-1}(\mathbf{b} - U\mathbf{x}^{(k)}).$$

However, by taking advantage of the triangular form of L_*, the elements of x(k+1) can be computed sequentially using forward substitution:

$$x_i^{(k+1)} = \frac{1}{a_{ii}}\left(b_i - \sum_{j<i} a_{ij}x_j^{(k+1)} - \sum_{j>i} a_{ij}x_j^{(k)}\right), \quad i, j = 1, 2, \ldots, n.$$

The procedure is generally continued until the changes made by an iteration are below some tolerance, such as a sufficiently small residual.

Theorem 3.4. *(convergence of the Gauss-Seidel Method) If A is diagonal dominant , then the Gauss-Seidel Method convergences.*

Proof. We want to show If A is diagonal dominant , then $\|T_{GS}\| < 1$, then Gauss-Seidel Method convergences. From the definition of T, we know that T for Gauss-Seidel Method is as follows

$$T_{GS} = -(L + D)^{-1}U.$$

Next, we will show $\|T_{GS}\| < 1$. Since A is diagonal dominant, so

$$|a_{ii}| \geq \sum_{j\neq i} |a_{ij}| + \delta = \sum_{j>i} |a_{ij}| + \sum_{j<i} |a_{ij}| + \delta.$$

So,

$$|a_{ii}| - \sum_{j<i} |a_{ij}| \geq \sum_{j>i} |a_{ij}| + \delta,$$

which implies

$$\gamma = max_i \left\{ \frac{\sum_{j>i} |a_{ij}|}{|a_{ii}| - \sum_{j<i} |a_{ij}|} \right\} \leq 1.$$

Now, we will show $\|T_{GS}\| < \gamma$. Let $x \in \mathbb{C}^n$ and $y = Tx$, i.e.

$$y = T_{GS}x = -(L+D)^{-1}Ux.$$

Let i_0 be the index such that $\|y\|_\infty = |y_{i_0}|$, then we have

$$|((L+D)y)_{i_0}| = |(Ux)_{i_0}| = |\sum_{j>i_0} a_{i_0 j} x_j| \leq \sum_{j>i_0} |a_{i_0 j}||x_j| \leq \sum_{j>i_0} |a_{i_0 j}| \|x\|_\infty .$$

Moreover

$$|((L+D)y)_{i_0}| = |\sum_{j<i_0} a_{i_0 j} y_j + a_{i_0 i_0} y_j| \geq |a_{i_0 i_0} y_j| - |\sum_{j<i_0} a_{i_0 j} y_j| = |a_{i_0 i_0}| \|y\|_\infty - |\sum_{j<i_0} a_{i_0 j} y_j| \geq |a_{i_0 i_0}| \|y\|_\infty - \sum_{j<i_0} |a_{i_0 j}| \|y\|_\infty .$$

Therefore, we have

$$|a_{i_0 i_0}| \|y\|_\infty - \sum_{j<i_0} |a_{i_0 j}| \|y\|_\infty \leq \sum_{j>i_0} |a_{i_0 j}| \|x\|_\infty ,$$

which implies

$$\|y\|_\infty \leq \frac{\sum_{j>i_0} |a_{i_0 j}|}{|a_{i_0 i_0}| - \sum_{j<i_0} |a_{i_0 j}|} \|x\|_\infty .$$

So,

$$\|T_{GS}x\|_\infty \leq \gamma \|x\|_\infty ,$$

which implies

$$\|T_{GS}\|_\infty \leq \gamma < 1.$$

□

3.3.3 Richardson Method

Definition 3.8. *(Richardson Method) Let*

$$A = L + D + U.$$

A Richardson Method scheme reads

$$I\left(\frac{x^{k+1} - x^k}{\omega}\right) + Ax^k = b.$$

i.e. $\alpha_k = \omega \neq 1, B = I$ *in the general iterative scheme.*

Definition 3.9. *(Error Transfer Operator for Gauss-Seidel Method) The error transfer operator for Gauss-Seidel Method is as follows*

$$T_{RC} = I - \omega(B)^{-1}A = I - \omega A.$$

Remark 3.3. *Richardson iteration is an iterative method for solving a system of linear equations. Richardson iteration was proposed by Lewis Richardson in his work dated 1910. It is similar to the Jacobi and Gauss-Seidel method. We seek the solution to a set of linear equations, expressed in matrix terms as*

$$Ax = b.$$

The Richardson iteration is

$$x^{(k+1)} = (I - \omega A)x^{(k)} + \omega b.$$

where α is a scalar parameter that has to be chosen such that the sequence $x^{(k)}$ converges.

It is easy to see that the method has the correct fixed points, because if it converges, then $x^{(k+1)} \approx x^{(k)}$ and $x^{(k)}$ has to approximate a solution of $Ax = b$.

Theorem 3.5. *(convergence of the Richardson Method) Let $A = A^* > 0$ (SPD). If $0 < \omega < \frac{2}{\lambda_{\max}}$, then the Richardson Method convergences. Moreover, the best acceleration parameter is given by*

$$\omega_{opt} = \frac{2}{\lambda_{\min} + \lambda_{\max}},$$

in which, similarly, $\lambda_{\min}$ is the smallest eigenvalue of $A^T A$.

Proof. 1. From the above lemma, we know that the error transform operator is as follows

$$T_{RC} = I - \omega(B)^{-1}A = I - \omega A.$$

Let $\lambda \in \sigma(A)$, then $\nu := 1 - \omega\lambda \in \sigma(T)$. From the sufficient and & necessary condition for convergence, we know if $\sigma(T) < 1$, then Richardson Method convergences, i.e.

$$|1 - \omega\lambda| < 1,$$

which implies

$$-1 < 1 - \omega\lambda_{\max} \leq 1 - \omega\lambda_{\min} < 1.$$

So, we get $-1 < 1 - \omega\lambda_{\max}$, i.e.

$$\omega < \frac{2}{\lambda_{\max}}.$$

2. The minimum is attachment at $|1 - \omega\lambda_{\max}| = |1 - \omega\lambda_{\min}|$(Figure.1), i.e.

$$\omega\lambda_{\max} - 1 = 1 - \omega\lambda_{\min}.$$

Therefore, we get

$$\omega_{opt} = \frac{2}{\lambda_{\min} + \lambda_{\max}}.$$

□

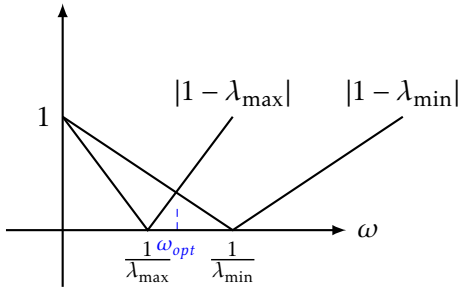

Figure 1: The curve of $\rho(T_{RC})$ as a function of ω

3.3.4 Successive Over Relaxation (SOR) Method

Definition 3.10. *(SOR Method) Let*

$$A = L + D + U.$$

A SOR Method scheme reads

$$(\omega L + D)\left(\frac{x^{k+1} - x^k}{\omega}\right) + Ax^k = b.$$

i.e. $\alpha_k = \omega \neq 1, B = \omega L + D$ *in the general iterative scheme.*

Remark 3.4. *For Gauss-seidel method, we have*

$$Lx^{k+1} + Dx^{k+1} + Ux^k = b.$$

If we relax the contribution of the diagonal part, i.e. let $\omega > 0$,

$$D = \omega^{-1}D + (1 - \omega^{-1})D,$$

and

$$Lx^{k+1} + \omega^{-1}Dx^{k+1} + (1 - \omega^{-1})Dx^k + Ux^k = b.$$

Then, we obtain

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

- $\omega = 1$-*Gauss-Seidel method,*
- $\omega < 1$-*Under relaxation method,*
- $\omega > 1$-*Over relaxation method.*

We can rewire the above formula to get the general form:

$$\begin{aligned}
(L + \omega^{-1}D)x^{k+1} + ((1 - \omega^{-1})D + U)x^k &= b. \\
(L + \omega^{-1}D)x^{k+1} + (D - \omega^{-1}D + U + L - L)x^k &= b \\
(L + \omega^{-1}D)x^{k+1} + (A - (L + \omega^{-1}D))x^k &= b \\
(L + \omega^{-1}D)(x^{k+1} - x^k) + Ax^k &= b \\
(\omega L + D)\frac{x^{k+1} - x^k}{\omega} + Ax^k &= b
\end{aligned}$$

Definition 3.11. *(Error Transfer Operator for Gauss-Seidel Method) The error transfer operator for SOR Method is as follows*

$$T_{SOR} = I - \alpha(B)^{-1}A = I - \omega(\omega L + D)^{-1}A = -(L + \omega^{-1}D)^{-1}((1-\omega^{-1})D + U).$$

Theorem 3.6. *(Necessary condition for convergence of the SOR Method) If SOR method convergences, then $0 < \omega < 2$.*

Proof. If SOR method convergences, then $\rho(T) < 1$, i.e $|\lambda| < 1$. Let λ_i are the roots of characteristic polynomial $X_T(\lambda) = det(\lambda I - T) = (-1)^n \Pi_{i=1}^n (\lambda - \lambda_i)$. Then,

$$X_T(0) = \Pi_{i=1}^n \lambda_i = det(T_{SOR}).$$

Since $\lambda_i < 1$, so $|det(T_{SOR})| < 1$. Since $T_{SOR} = -(\omega L + D)^{-1}((1-\omega^{-1})D + U)$, then

$$\begin{aligned} det(T_{SOR}) &= det((L+\omega^{-1}D)^{-1})det((1-\omega^{-1})D+U) \\ &= \frac{det((1-\omega^{-1})D+U)}{det(L+\omega^{-1}D)} = \frac{det((1-\omega^{-1})D)}{det(\omega^{-1}D)} = \frac{\Pi_{i=1}^n (1-\omega^{-1})a_{ii}}{\Pi_{i=1}^n \omega^{-1}a_{ii}} \\ &= \frac{(1-\omega^{-1})^n}{\omega^{-n}} = |\omega - 1|^n < 1 \end{aligned}$$

Therefore, $|\omega - 1| < 1$, so $0 < \omega < 2$. □

Theorem 3.7. *(convergence of the SOR Method for SPD) If $A = A^*$, and $0 < \omega < 2$, then SOR converges.*

Proof. Since

$$T_{SOR} = -(L+\omega^{-1}D)^{-1}((1-\omega^{-1})D + U) = (L+\omega^{-1}D)^{-1}((\omega^{-1}-1)D - U).$$

Let $Q = L + \omega^{-1}D$, then

$$I - T_{SOR} = Q^{-1}A.$$

Let (λ, x) be the eigenpair of T, i.e. $Tx = \lambda x$ and $y = (I - T_{SOR})x = (1-\lambda)x$. So, we have

$$y = Q^{-1}Ax, \text{ or } Qy = Ax.$$

Moreover,

$$(Q-A)y = Qy - Ay = Ax - Ay = A(x-y) = A(x-(I-T)x) = ATx = \lambda Ax.$$

So, we have

$$\begin{aligned} (Qy, y) &= (Ax, y) = (Ax, (1-\lambda)x) = (1-\bar{\lambda})(Ax, x). \\ (y, (Q-A)y) &= (y, \lambda Ax) = \bar{\lambda}(y, Ax) = \bar{\lambda}((1-\lambda)x, Ax) = \bar{\lambda}(1-\bar{\lambda})(x, Ax) = \bar{\lambda}(1-\bar{\lambda})(Ax, x). \end{aligned}$$

Plus the above equation together, then

$$(Qy, y) + (y, (Q-A)y) = (1-\bar{\lambda})(Ax, x) + \bar{\lambda}(1-\bar{\lambda})(Ax, x) = (1-|\lambda|^2)(Ax, x).$$

while

$$\begin{aligned} (Qy, y) + (y, (Q-A)y) &= ((L+\omega^{-1}D)y, y) + (y, (L+\omega^{-1}D-A)y) \\ &= (Ly, y) + (\omega^{-1}Dy, y) + (y, \omega^{-1}Dy) - (y, Dy) - (y, Uy) \\ &= (2\omega^{-1} - 1)(Dy, y). (\text{since} A = A^*, \text{so}, L = U) \end{aligned}$$

So, we get

$$(2\omega^{-1}-1)(Dy,y)=(1-|\lambda|^2)(Ax,x).$$

Since $0<\omega<2$, $(Dy,y)>0$ and $(Ax,x)>0$, so we have

$$(1-|\lambda|^2)>0.$$

Then, we have $|\lambda|<1$. □

3.4 Convergence in energy norm for steady cases

From now on, $A=A^*>0$.

Definition 3.12. *(Energy norm w.r.t A) The Energy norm associated with A is*

$$\|x\|_A=(Ax,x);$$

Now, we will consider the convergence in energy norm of stationary scheme,

$$B\left(\frac{x^{k+1}-x^k}{\alpha}\right)+Ax^k=b.$$

Theorem 3.8. *(convergence in energy norm) If $Q=B-\frac{\alpha}{2}A>0$, then $\left\|e^k\right\|_A\to 0$.*

Proof. Let $e^k=x^k-x$. Since

$$B\left(\frac{x^{k+1}-x^k}{\alpha}\right)+Ax^k=b=Ax.$$

so, we get

$$B\left(\frac{e^{k+1}-e^k}{\alpha}\right)+Ae^k=0.$$

Let $v^{k+1}=e^{k+1}-e^k$, then

$$\frac{1}{\alpha}Bv^{k+1}+Ae^k=0.$$

Then take the inner product of both sides with v^{k+1},

$$\frac{1}{\alpha}(Bv^{k+1},v^{k+1})+(Ae^k,v^{k+1})=0.$$

Since

$$e^k=\frac{1}{2}(e^{k+1}+e^k)-\frac{1}{2}(e^{k+1}-e^k)=\frac{1}{2}(e^{k+1}+e^k)-\frac{1}{2}v^{k+1}.$$

Therefore,

$$\begin{aligned}
0 &= \frac{1}{\alpha}(Bv^{k+1},v^{k+1})+(Ae^k,v^{k+1})\\
&= \frac{1}{\alpha}(Bv^{k+1},v^{k+1})+\frac{1}{2}(A(e^{k+1}+e^k),v^{k+1})-\frac{1}{2}(Av^{k+1},v^{k+1})\\
&= \frac{1}{\alpha}((B-\frac{\alpha}{2}A)v^{k+1},v^{k+1})+\frac{1}{2}(A(e^{k+1}+e^k),v^{k+1})\\
&= \frac{1}{\alpha}((B-\frac{\alpha}{2}A)v^{k+1},v^{k+1})+\frac{1}{2}(\left\|e^{k+1}\right\|_A^2-\left\|e^k\right\|_A^2)
\end{aligned}$$

By assumption, $Q = B - \frac{\alpha}{2}A > 0$, i.e. there exists $m > 0$, s.t.

$$(Qy, y) \geq m\|y\|_2^2.$$

Therefore,

$$\frac{m}{\alpha}\|v^{k+1}\|_2^2 + \frac{1}{2}(\|e^{k+1}\|_A^2 - \|e^k\|_A^2) \leq 0.$$

i.e.

$$\frac{2m}{\alpha}\|v^{k+1}\|_2^2 + \|e^{k+1}\|_A^2 \leq \|e^k\|_A^2.$$

Hence

$$\|e^{k+1}\|_A^2 \leq \|e^k\|_A^2.$$

and

$$\|e^{k+1}\|_A^2 \to 0.$$

□

3.5 Dynamic cases iterative method

In this subsection, we will consider the following dynamic iterative method

$$B_k\left(\frac{x^{k+1} - x^k}{\alpha_k}\right) + Ax^k = b.$$

where B_k and α_k are dependent on the k.

3.5.1 Chebyshev iterative Method

Definition 3.13. *(Chebyshev iterative Method) Chebyshev iterative Method is going to choose $\alpha_1, \alpha_2, \cdots, \alpha_k$, s.t. $\|e^k\|_2$ is minimal for*

$$\left(\frac{x^{k+1} - x^k}{\alpha_{k+1}}\right) + Ax^k = b.$$

Theorem 3.9. *(convergence of Chebyshev iterative Method) If $A = A^* > 0$, then for a given n, $\|e^k\|$ is minimized by choosing*

$$\alpha_k = \frac{\alpha_0}{1 + \rho_0 t_k}, t = 1, \cdots, n.$$

Where

$$\alpha_0 = \frac{2}{\lambda_{\min} + \lambda_{\max}}, \quad \rho_0 = \frac{\kappa_2(A) - 1}{\kappa_2(A) + 1}, \quad t_k = \cos\left(\frac{(2k+1) * 2\pi}{2n}\right).$$

Moreover, we have

$$\|e^k\|_2 \leq 2\frac{\rho_1^k}{1 + \rho_1^{2k}}\|e^0\|_2, \quad \text{where } \rho_1 = \frac{\sqrt{\kappa_2(A)} - 1}{\sqrt{\kappa_2(A)} + 1}.$$

3.5.2 Minimal residuals Method

Definition 3.14. *(Minimal residuals Method) Minimal residuals iterative Method is going to choose $\alpha_1, \alpha_2, \cdots, \alpha_k$, s.t. the residuals $r^k = b - Ax^k$ is minimal for*

$$\left(\frac{x^{k+1} - x^k}{\alpha_{k+1}}\right) + Ax^k = b.$$

Theorem 3.10. *(optimal α_{k+1} of minimal residuals iterative Method) The optimal α_{k+1} of minimal residuals iterative Method is as follows*

$$\alpha_{k+1} = \frac{(r^k, Ar^k)}{\left\|Ar^k\right\|_2^2}.$$

Proof. From the iterative scheme

$$\left(\frac{x^{k+1} - x^k}{\alpha_{k+1}}\right) + Ax^k = b,$$

we get

$$x^{k+1} = x^k + \alpha_{k+1} r^k.$$

By multiplying $-A$ and add b to both side of the above equation, we have

$$r^{k+1} = r^k - \alpha_{k+1} Ar^k.$$

Therefore,

$$\begin{aligned} \left\|r^{k+1}\right\|_2^2 &= (r^k - \alpha_{k+1} Ar^k, r^k - \alpha_{k+1} Ar^k) \\ &= \left\|r^k\right\|_2^2 - 2\alpha_{k+1}(r^k, Ar^k) + \alpha_{k+1}^2 \left\|Ar^k\right\|_2^2. \end{aligned}$$

When α_{k+1} minimize the residuals, the

$$(\left\|r^{k+1}\right\|_2^2)' = -2(r^k, Ar^k) + 2\alpha_{k+1} \left\|Ar^k\right\|_2^2 = 0, i.e.$$

$$\alpha_{k+1} = \frac{(r^k, Ar^k)}{\left\|Ar^k\right\|_2^2}.$$

□

Corollary 3.1. *The residual r^{k+1} of minimal residuals iterative Method is orthogonal to residual r^k in A-norm.*

Proof.

$$(Ar^{k+1}, r^k) = (r^{k+1}, Ar^k) = (r^k - \alpha_{k+1} Ar^k, Ar^k) = (r^k, Ar^k) - \alpha_{k+1}(Ar^k, Ar^k) = 0.$$

□

Algorithm 3.1. *(Minimal residuals method algorithm)*

- x^0
- *compute* $r^k = b - Ax^k$
- *compute* $\alpha_{k+1} = \frac{(r^k, Ar^k)}{\|Ar^k\|_2^2}$
- *compute* $x^{k+1} = x^k + \alpha_{k+1} r^k$

Theorem 3.11. *(convergence of minimal residuals iterative Method)* *The minimal residuals iterative Method converges for any* x^0 *and*

$$\left\|Ae^k\right\|_2 \leq \rho_0^n \left\|Ae^0\right\|_2, \text{ with } \rho_0 = \frac{\kappa_2(A)-1}{\kappa_2(A)+1}.$$

Proof. Since the choice

$$\alpha_{k+1} = \frac{(r^k, Ar^k)}{\left\|Ar^k\right\|_2^2}.$$

minimizes the $\left\|r^{k+1}\right\|$. Consequently, choosing

$$\alpha_{k+1} = \alpha_0 = \frac{1}{\lambda_{\max} + \lambda_{\min}},$$

we get

$$\rho_0 = \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} = \frac{\frac{\lambda_{\max}}{\lambda_{\min}} - 1}{\frac{\lambda_{\max}}{\lambda_{\min}} + 1} = \frac{\|A\|_2 \left\|A^{-1}\right\|_2 - 1}{\|A\|_2 \left\|A^{-1}\right\|_2 + 1} = \frac{\kappa_2(A)-1}{\kappa_2(A)+1}.$$

Moreover, since

$$r^{k+1} = r^k - \alpha_{k+1} Ar^k = (I - \alpha_{k+1} A) r^k,$$

then

$$\left\|r^{k+1}\right\|_2 \leq \left\|I - \alpha_{k+1} A\right\|_2 \left\|r^k\right\|_2 = \rho(T) \leq \rho_0 \left\|r^k\right\|_2.$$

Since

$$Ae^k = A(x - x^k) = Ax - Ax^k = b - Ax^k = r^k,$$

so,

$$\left\|Ae^{k+1}\right\|_2 = \left\|r^{k+1}\right\|_2 \leq \left\|I - \alpha_{k+1} A\right\|_2 \left\|r^k\right\|_2 = \rho(T) \leq \rho_0 \left\|r^k\right\|_2 \leq \rho_0^n \left\|Ae^0\right\|_2.$$

□

3.5.3 Minimal correction iterative method

Definition 3.15. *(Minimal correction Method)* *Minimal correction iterative Method is going to choose* $\alpha_1, \alpha_2, \cdots, \alpha_k$, *s.t. the correction* $\left\|w^{k+1}\right\|_B (w^k = B^{-1}(b - Ax^k) = B^{-1} r^k, A = A^* > 0, B = B^* > 0)$ *is minimal for*

$$B\left(\frac{x^{k+1} - x^k}{\alpha_{k+1}}\right) + Ax^k = b.$$

Theorem 3.12. *(optimal α_{k+1} of minimal correction iterative Method) The optimal α_{k+1} of minimal correction iterative Method is as follows*

$$\alpha_{k+1} = \frac{(w^k, Aw^k)}{(B^{-1}Aw^k, Aw^k)} = \frac{\left\|w^k\right\|_A}{\left\|Aw^k\right\|_{B^{-1}}}.$$

Proof. From the iterative scheme

$$B\left(\frac{x^{k+1}-x^k}{\alpha_{k+1}}\right) + Ax^k = b,$$

we get

$$x^{k+1} = x^k + \alpha_{k+1}B^{-1}r^k.$$

By multiplying $-A$ and add b to both side of the above equation, we have

$$r^{k+1} = r^k - \alpha_{k+1}AB^{-1}r^k.$$

Since, $w^k = B^{-1}(b - Ax^k) = B^{-1}r^k$, $A = A^* > 0, B = B^* > 0$ Therefore,

$$\begin{aligned}
\left\|w^{k+1}\right\|_B^2 &= (Bw^{k+1}, w^{k+1}) = (BB^{-1}r^{k+1}, B^{-1}r^{k+1}) = (r^{k+1}, B^{-1}r^{k+1}) \\
&= (r^k - \alpha_{k+1}AB^{-1}r^k, B^{-1}r^k - \alpha_{k+1}B^{-1}AB^{-1}r^k) \\
&= (r^k, B^{-1}r^k) - \alpha_{k+1}(r^k, B^{-1}AB^{-1}r^k) - \alpha_{k+1}(AB^{-1}r^k, B^{-1}r^k) - \alpha_{k+1}^2(AB^{-1}r^k, B^{-1}AB^{-1}r^k) \\
&= (r^k, B^{-1}r^k) - 2\alpha_{k+1}(B^{-1}r^k, AB^{-1}r^k) + \alpha_{k+1}^2(B^{-1}AB^{-1}r^k, AB^{-1}r^k) \\
&= (r^k, w^k) - 2\alpha_{k+1}(w^k, Aw^k) + \alpha_{k+1}^2(B^{-1}Aw^k, Aw^k)
\end{aligned}$$

When α_{k+1} minimize the residuals, the

$$(\left\|w^{k+1}\right\|_B^2)' = -2(w^k, Aw^k) + 2\alpha_{k+1}(B^{-1}Aw^k, Aw^k) = 0, i.e.$$

$$\alpha_{k+1} = \frac{(w^k, Aw^k)}{(B^{-1}Aw^k, Aw^k)}.$$

□

Remark 3.5. *Most of time, it's not easy to compute $\|\cdot\|_A, \|\cdot\|_{B^{-1}}$. We will use the following alternative way to implement the algorithm. let $v^k = B^{\frac{1}{2}}w^k$, then from the iterative scheme*

$$B\left(\frac{x^{k+1}-x^k}{\alpha_{k+1}}\right) + Ax^k = b,$$

Multiplying by B^{-1} on both side of the above equation yields

$$\left(\frac{x^{k+1}-x^k}{\alpha_{k+1}}\right) + B^{-1}Ax^k = B^{-1}b.$$

Then, Multiplying by $-A$ on both side of the above equation yields

$$\left(\frac{-Ax^{k+1}+Ax^k}{\alpha_{k+1}}\right) - AB^{-1}Ax^k = -AB^{-1}b.$$

therefore

$$\left(\frac{b-Ax^{k+1}-(b-Ax^k)}{\alpha_{k+1}}\right)+AB^{-1}(b-Ax^k)=0,$$

i.e.

$$\left(\frac{r^{k+1}-r^k}{\alpha_{k+1}}\right)+AB^{-1}r^k=0.$$

By using the identity $B^{-1}r^k=w^k$, *we get*

$$B\left(\frac{w^{k+1}-w^k}{\alpha_{k+1}}\right)+Aw^k=0.$$

Then, we have

$$B^{\frac{1}{2}}B^{\frac{1}{2}}\left(\frac{w^{k+1}-w^k}{\alpha_{k+1}}\right)+AB^{-\frac{1}{2}}B^{\frac{1}{2}}w^k=b.$$

Multiplying by $B^{-\frac{1}{2}}$ *on both side of the above equation yields*

$$B^{\frac{1}{2}}\left(\frac{w^{k+1}-w^k}{\alpha_{k+1}}\right)+B^{-\frac{1}{2}}AB^{-\frac{1}{2}}B^{\frac{1}{2}}w^k=B^{-\frac{1}{2}}b,$$

i.e.

$$B\left(\frac{v^{k+1}-v^k}{\alpha_{k+1}}\right)+B^{-\frac{1}{2}}AB^{-\frac{1}{2}}v^k=0.$$

Since $B^{-\frac{1}{2}}AB^{-\frac{1}{2}}>0$, *then we minimize* $\left\|v^{k+1}\right\|_2$ *instead of* $\left\|w^{k+1}\right\|_B$. *But*

$$\left\|w^{k+1}\right\|_B^2=(Bw^{k+1},w^{k+1})=(B^{\frac{1}{2}}B^{\frac{1}{2}}w^{k+1},w^{k+1})=(B^{\frac{1}{2}}w^{k+1},B^{\frac{1}{2}}w^{k+1})=\left\|v^{k+1}\right\|_2^2.$$

Theorem 3.13. *(convergence of minimal correction iterative Method) The minimal correction iterative Method converges for any* x^0 *and*

$$\left\|Ae^k\right\|_{B^{-1}}\leq\rho_0^n\left\|Ae^0\right\|_{B^{-1}},\ \ with\ \ \rho_0=\frac{\kappa_2(B^{-1}A)-1}{\kappa_2(B^{-1}A)+1}.$$

Proof. Same as convergence of minimal residuals iterative Method. □

Algorithm 3.2. *(Minimal correction method algorithm)*

- x^0
- *compute* $w^k=B^{-1}(b-Ax^k)$
- *compute* $\alpha_{k+1}=\frac{(w^k,Aw^k)}{(B^{-1}Aw^k,Aw^k)}$
- *compute* $x^{k+1}=x^k+\alpha_{k+1}w^k$

3.5.4 Steepest Descent Method

Definition 3.16. *(Steepest Descent Method) Steepest Descent iterative Method is going to choose $\alpha_1, \alpha_2, \cdots, \alpha_k$, s.t. the error $\left\|e^{k+1}\right\|_A$ is minimal for*

$$\left(\frac{x^{k+1}-x^k}{\alpha_{k+1}}\right)+Ax^k=b.$$

Theorem 3.14. *(optimal α_{k+1} of Steepest Descent iterative Method) The optimal α_{k+1} of Steepest Descent iterative Method is as follows*

$$\alpha_{k+1}=\frac{\left\|Ae^k\right\|_2^2}{\left\|Ae^k\right\|_A^2}=\frac{\left\|r^k\right\|_2^2}{\left\|r^k\right\|_A^2}.$$

Proof. From the iterative scheme

$$\left(\frac{x^{k+1}-x^k}{\alpha_{k+1}}\right)+Ax^k=b=Ax,$$

we get

$$e^{k+1}=e^k+\alpha_{k+1}Ae^k.$$

Therefore

$$\begin{aligned}\left\|e^{k+1}\right\|_A^2 &= (Ae^{k+1},e^{k+1})\\ &= (Ae^k+\alpha_{k+1}A^2e^k,e^k+\alpha_{k+1}Ae^k)\\ &= \left\|e^k\right\|_A^2-2\alpha_{k+1}\left\|Ae^k\right\|_2^2+\alpha_{k+1}^2\left\|Ae^k\right\|_A^2\end{aligned}$$

When α_{k+1} minimize the residuals, the

$$(\left\|e^{k+1}\right\|_2^2)'=-2\left\|Ae^k\right\|_2^2+2\alpha_{k+1}\left\|Ae^k\right\|_A^2=0, i.e.$$

$$\alpha_{k+1}=\frac{\left\|Ae^k\right\|_2^2}{\left\|Ae^k\right\|_A^2}=\frac{\left\|r^k\right\|_2^2}{\left\|r^k\right\|_A^2}.$$

The last step, we use the fact $Ae^k=r^k$. □

Theorem 3.15. *(convergence of Steepest Descent iterative Method) The Steepest Descent iterative Method converges for any x^0 ($A=A^*>0, B=B^*>0$) and*

$$\left\|e^k\right\|_A\le\rho_0^n\left\|e^0\right\|_A,\ \ with\ \ \rho_0=\frac{\kappa_2(A)-1}{\kappa_2(A)+1}.$$

Proof. Same as convergence of minimal residuals iterative Method. □

3.5.5 Conjugate Gradients Method

Definition 3.17. *(Conjugate Gradients Method) Conjugate Gradients Method iterative Method is a three-layer iterative method which is going to choose $\alpha_1,\alpha_2,\cdots,\alpha_k$ and $\tau_1,\tau_2,\cdots,\tau_k$, s.t. the error $\left\|e^{k+1}\right\|_A$ is minimal for*

$$B\frac{(x^{k+1}-x^k)+(1-\alpha_{k+1})(x^k-x^{k-1})}{\alpha_{k+1}\tau_{k+1}}+Ax^k=b.$$

3.5.6 Another look at Conjugate Gradients Method

If A is SPD, we know that solving $Ax=b$ is equivalent to minimize the following quadratic functional

$$\Phi(x)=\frac{1}{2}(Ax,x)-(f,x).$$

In fact, the minimum value of Φ is $-\frac{1}{2}(A^{-1}f,f)$ at $x=A^{-1}f$ and the residual r^k is the negative gradient of Φ at x^k, i.e.

$$r^k=-\nabla\Phi(x^k).$$

- Richardson method is always using the increment along the negative gradient of Φ to correct the result, i.e.

$$x^{k+1}=x^k+\alpha_k r^k.$$

- Conjugate Gradients Method is using the increment along the direction p^k which is not parallel to the gradient of Φ to correct the result.

Definition 3.18. *(A-Conjugate) The direction $\{p^k\}$ is call A-Conjugate, if $(p^j,Ap^k)=0$ when $j\neq k$. In particular,*

$$(p^{k+1},Ap^k)=0,\ \ \forall k\in\mathbb{N}.$$

Let $p^0,p^1,\cdots,p^m$ be the linearly independent series and x^0 be the initial guess, then we can construct the following series

$$x^{k+1}=x^k+\alpha_k p^k,\ \ 0\le k\le m.$$

where α_k is nonnegative. And then the minimum functional $\Phi(x)$ of x^{k+1} on $k+1$ dimension hyperplane is

$$x=x^0+\sum_{j=0}^{k}\gamma_j p^j,\ \ \gamma_j\in\mathbb{R}$$

if and only if p^j is A-Conjugate and

$$\alpha_k=\frac{(r^k,p^k)}{(p^{k+1},Ap^k)}.$$

Algorithm 3.3. *(Conjugate Gradients method algorithm)*

- x^0
- *compute* $r^0 = f - Ax^0$ *and* $p^0 = r^0$
- *compute* $\alpha_k = \frac{(r^k, p^k)}{(p^k, Ap^k)} = \frac{\|r^k\|_2^2}{(p^k, Ap^k)}$
- *compute* $x^{k+1} = x^k + \alpha_k p^k$
- *compute* $r^{k+1} = r^k - \alpha_k A p^k$
- *compute* $\beta_{k+1} = -\frac{(r^{k+1}, Ap^k)}{(p^k, Ap^k)} = -\frac{\|r^{k+1}\|_2^2}{(p^k, Ap^k)}$
- *compute* $x^{k+1} = x^k + \beta_{k+1} p^k$

Properties 3.2. *(properties of $\{p^k\}$ and $\{r^k\}$) the $\{p^k\}$ and $\{r^k\}$ come from the Conjugate Gradients method have the following properties:*

- $(p^j, r^j) = 0, \quad 0 \le i < j \le k$
- $(p^i, Ap^j) = 0, \quad i \ne j \quad 0 \le i, j \le k$
- $(r^i, r^j) = 0, \quad i \ne j \quad 0 \le i, j \le k$

Theorem 3.16. *(convergence of Conjugate Gradients iterative Method) The Conjugate Gradients iterative Method converges for any x^0 ($A = A^* > 0, B = B^* > 0$) and*

$$\left\|e^k\right\|_A \le 2\rho_0^n \left\|e^0\right\|_A, \quad with \quad \rho_0 = \frac{\sqrt{\kappa_2(A)} - 1}{\sqrt{\kappa_2(A)} + 1}.$$

Definition 3.19. *(Krylov subspace) In linear algebra, the order-k Krylov subspace generated by an n-by-n matrix A and a vector b of dimension n is the linear subspace spanned by the images of b under the first k-1 powers of A (starting from $A^0 = I$), that is,*

$$\mathcal{K}_k(A, b) = \text{span}\{b, Ab, A^2 b, \ldots, A^{k-1} b\}.$$

Theorem 3.17. *(Conjugate Gradients iterative Method in Krylov subspace) For Conjugate Gradients iterative Method, we have*

$$span\{r^0, r^1, \cdots, r^k\} = span\{p^0, p^1, \cdots, p^k\} = \mathcal{K}_{k+1}(A, r^0).$$

3.6 Problems

Problem 3.1. *(Prelim Jan. 2011#1) Consider a linear system $Ax = b$ with $A \in \mathbb{R}^{n\times n}$. Richardson's method is an iterative method*

$$Mx^{k+1} = Nx^k + b$$

with $M = \frac{1}{w}, N = M - A = \frac{1}{w}I - A$, where w is a damping factor chosen to make M approximate A as well as possible. Suppose A is positive definite and $w > 0$. Let λ_1 and λ_n denote the smallest and largest eigenvalue of A.

1. *Prove that Richardson's method converges if and only if $w < \frac{2}{\lambda_n}$.*
2. *Prove that the optimal value of w is $w_0 = \frac{2}{\lambda_1+\lambda_n}$.*

Solution. 1. Since $M = \frac{1}{w}, N = M - A = \frac{1}{w}I - A$, then we have

$$x^{k+1} = (I - wA)x^k + bw.$$

So $T_R = I - wA$, From the sufficient and & necessary condition for convergence, we should have $\rho(T_R) < 1$. Since λ_i are the eigenvalue of A, then we have $1 - \lambda_i w$ are the eigenvalues of T_R. Hence Richardson's method converges if and only if $|1 - \lambda_i w| < 1$, i.e

$$-1 < 1 - \lambda_n w < \cdots < 1 - \lambda_1 w < 1,$$

i.e. $w < \frac{2}{\lambda_n}$.

2. the minimal attaches at $|1 - \lambda_n w| = |1 - \lambda_1 w|$ (Figure. B2), i.e

$$\lambda_n w - 1 = 1 - \lambda_1 w,$$

i,e

$$w_0 = \frac{2}{\lambda_1 + \lambda_n}.$$

◀

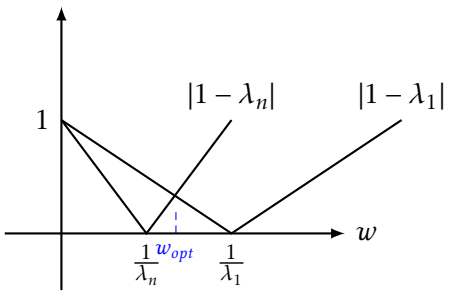

Figure 2: The curve of $\rho(T_R)$ as a function of w

Problem 3.2. *(Prelim Aug. 2010#3) Suppose that $A \in \mathbb{R}^{n\times n}$ is SPD and $b \in \mathbb{R}^n$ is given. Then n^{th} Krylov subspace us defined as*

$$\mathcal{K}_n := \left\langle b, Ab, A^2b, \cdots, A^{k-1}b \right\rangle.$$

Let $\left\{x_j\right\}_{j=0}^{n-1}, x_0 = 0$, denote the sequence of vectors generated by the conjugate gradient algorithm. Prove that if the method has not already converged after $n-1$ iterations, i.e. $r_{n-1} = b - Ax_{n-1} \neq 0$, then the n^{th} iterate x_n us the unique vector in $\mathcal{K}_n$ that minimizes

$$\phi(y) = \left\|x_* - y\right\|_A^2,$$

where $x_* = A^{-1}b$.

Solution. ◀

Problem 3.3. *(Prelim Jan. 2011#1)*

Solution. ◀

4 Eigenvalue Problems

Definition 4.1. *(Geršchgorin disks) Let $A \in \mathbb{C}^{n\times n}$, the Geršchgorin disks of A are*

$$D_i = \{\xi \in \mathbb{C} : |\xi - a_{ii}| < \mathbb{R}_i\} \quad \textit{where} \quad \mathbb{R}_i = \sum_{i\neq j} |a_{ij}|.$$

Theorem 4.1. *Every eigenvalue of A lies within at least one of the Geršchgorin discs D_i*

Proof. Let λ be an eigenvalue of A and let $x = (x_j)$ be a corresponding eigenvector. Let $i \in \{1, \cdots, n\}$ be chosen so that $|x_i| = \max_j |x_j|$. (That is to say, choose i so that x_i is the largest (in absolute value) number in the vector x) Then $|x_i| > 0$, otherwise x = 0. Since x is an eigenvector, $Ax = \lambda x$, and thus:

$$\sum_j a_{ij}x_j = \lambda x_i \quad \forall i \in \{1, \ldots, n\}.$$

So, splitting the sum, we get

$$\sum_{j\neq i} a_{ij}x_j = \lambda x_i - a_{ii}x_i.$$

We may then divide both sides by x_i (choosing i as we explained, we can be sure that $xi \neq 0$) and take the absolute value to obtain

$$|\lambda - a_{ii}| = \left|\frac{\sum_{j\neq i} a_{ij}x_j}{x_i}\right| \leq \sum_{j\neq i}\left|\frac{a_{ij}x_j}{x_i}\right| \leq \sum_{j\neq i}|a_{ij}| = R_i$$

where the last inequality is valid because

$$\left|\frac{x_j}{x_i}\right| \leq 1 \quad \text{for } j \neq i.$$

□

Corollary 4.1. *The eigenvalues of A must also lie within the Geršchgorin discs D_i corresponding to the columns of A.*

Proof. Apply the Theorem to A^T. □

Definition 4.2. *(Reyleigh Quotient) Let $A \in \mathbb{R}^{n\times n}$, $x \in \mathbb{R}^n$. The Reyleigh Quotient is*

$$\mathcal{R}(x) = \frac{(Ax, x)}{(x, x)}.$$

Remark 4.1. *If x is an eigenvector of A, then $Ax = \lambda x$ and*

$$\mathcal{R}(x) = \frac{(Ax, x)}{(x, x)} = \lambda.$$

Properties 4.1. *(properties of Reyleigh Quotient) Reyleigh Quotient has the following properties:*

1.
$$\nabla \mathcal{R}(x) = \frac{2}{(x,x)}[Ax - \mathcal{R}(x)x]$$

2. $\mathcal{R}(x)$ *minimizes*
$$f(\alpha) = \|Ax - \alpha x\|_2 .$$

Proof. 1. From the definition of the gradient, we have
$$\nabla \mathcal{R}(x) = \left[\frac{\partial r(x)}{\partial x_1}, \frac{\partial r(x)}{\partial x_2}, \cdots, \frac{\partial r(x)}{\partial x_n}\right].$$
By using the quotient rule, we have
$$\frac{\partial r(x)}{\partial x_i} = \frac{\partial}{\partial x_i}\left(\frac{(Ax,x)}{(x,x)}\right) = \frac{\partial}{\partial x_i}\left(\frac{x^T Ax}{x^T x}\right) = \frac{\frac{\partial}{\partial x_i}\left(x^T Ax\right)x^T x - x^T Ax\frac{\partial}{\partial x_i}\left(x^T x\right)}{\left(x^T x\right)^2},$$
where
$$\begin{aligned}
\frac{\partial}{\partial x_i}\left(x^T Ax\right) &= \frac{\partial}{\partial x_i}\left(x^T\right)Ax + x^T\frac{\partial}{\partial x_i}(Ax) \\
&= [0,\cdots,0,\underset{i}{1},0,\cdots,0]Ax + x^T A\begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} i \\
&= (Ax)_i + (Ax)_i = 2(Ax)_i .
\end{aligned}$$
Similarly,
$$\begin{aligned}
\frac{\partial}{\partial x_i}\left(x^T x\right) &= \frac{\partial}{\partial x_i}\left(x^T\right)x + x^T\frac{\partial}{\partial x_i}(x) \\
&= [0,\cdots,0,\underset{i}{1},0,\cdots,0]x + x^T\begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} i \\
&= 2x_i .
\end{aligned}$$
Therefore, we have
$$\begin{aligned}
\frac{\partial r(x)}{\partial x_i} &= \frac{2(Ax)_i}{x^T x} - \frac{x^T Ax 2x_i}{\left(x^T x\right)^2} \\
&= \frac{2}{x^T x}\left((Ax)_i - \mathcal{R}(x)x_i\right).
\end{aligned}$$

Hence

$$\nabla \mathcal{R}(x) = \frac{2}{x^T x}(Ax - \mathcal{R}(x)x) = \frac{2}{(x,x)}(Ax - \mathcal{R}(x)x).$$

2. let

$$g(\alpha) = \|Ax - \alpha x\|_2^2.$$

Then,

$$g(\alpha) = (Ax - \alpha x, Ax - \alpha x) = (Ax, Ax) - 2\alpha(Ax, x) + \alpha^2(x, x),$$

and

$$g'(\alpha) = -2(Ax, x) + 2\alpha(x, x),$$

when $\mathcal{R}(x)$ minimizes

$$f(\alpha) = \|Ax - \alpha x\|_2$$

, then $g'(\alpha) = 0$, i.e.

$$\alpha = \frac{(Ax, x)}{(x, x)} = \mathcal{R}(x).$$

□

4.1 Schur algorithm

Algorithm 4.1. *(Schur algorithm)*

- $A^0 = A = Q * UQ$
- *compute* $A_k = Q_k^{-1} A_{k-1} Q_k$

4.2 QR algorithm

Algorithm 4.2. *(QR algorithm)*

- $A^0 = A$
- *compute* $Q^k R^k = A^{k-1}$
- *compute* $A^k = R^k Q^k$

Properties 4.2. *(properties of QR algorithm) QR algorithm has the following properties:*

1. A^k *is similar to* A^{k-1}
2. $A^{k-1} = \left(A^{k-1}\right)^*$ *and* $A^k = \left(A^k\right)^*$
3. *If* A^{k-1} *is tridiagonal, then* A^k *is tridiagonal.*

Proof. 1. Since $Q^k R^k = A^{k-1}$, so $R^k = \left(Q^k\right)^{-1} A^{k-1}$ and $A^k = R^k Q^k = \left(Q^k\right)^{-1} A^{k-1} Q^k$.

2. Since Q is unitary, so $Q^* = Q^{-1}$ and $A = A^*$, so

$$\left(A^k\right)^* = \left(\left(Q^k\right)^{-1} A^{k-1} Q^k\right)^* = \left(Q^k\right)^* \left(A^{k-1}\right)^* \left(\left(Q^k\right)^{-1}\right)^* = \left(Q^k\right)^{-1} \left(A^{k-1}\right)^* \left(Q^k\right) = \left(Q^k\right)^{-1} A^{k-1} \left(Q^k\right) = A^k.$$

   Similarly,

$$\left(A^{k-1}\right)^* = \left(Q^k A^k \left(Q^k\right)^{-1}\right)^* = Q^k A^k \left(Q^k\right)^{-1} = A^{k-1}.$$

3. since A^k is similar to A^{k-1}.

□

4.3 Power iteration algorithm

Algorithm 4.3. *(Power iteration algorithm)*

- v^0: *an arbitrary nonzero vector*
- *compute* $v^k = Av^{k-1}$

Remark 4.2. *This algorithm generates a sequence of vectors*

$$v^0, Av^0, A^2v^0, A^3v^0, \cdots.$$

If we want to prove that this sequence converges to an eigenvector of A, the matrix needs to be such that it has a unique largest eigenvalue λ_1,

$$|\lambda_1| > |\lambda_2| \geq \cdots |\lambda_m| \geq 0.$$

There is another technical assumption. The initial vector v^0 *needs to be chosen such that* $q_1^T v^0 \neq 0$ *. Otherwise, if* v^0 *is completely perpendicular to the eigenvector* q_1*, the algorithm will not converge.*

Algorithm 4.4. *(improved Power iteration algorithm)*

- v^0: *an arbitrary nonzero vector with* $\left\|v^0\right\|_2 = 1$
- *compute* $w^k = Av^{k-1}$
- *compute* $v^k = \frac{w^k}{\|w^k\|_2}$
- *compute* $\lambda^k = \mathcal{R}(v^k)$

Theorem 4.2. *(Convergence of power algorithm) If* $A = A^*$, $q_1^T v^0 \neq 0$ *and* $|\lambda_1| > |\lambda_2| \geq \cdots |\lambda_m| \geq 0$, *then the convergence to the eigenvector of improved Power iteration algorithm is linear, while the convergence to the eigenvalue is still quadratic, i.e.*

$$\left\|v^k - (\pm q_1)\right\|_2 = \mathcal{O}\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right)$$

$$\left\|\lambda^k - \lambda_1\right\|_2 = \mathcal{O}\left(\left|\frac{\lambda_2}{\lambda_1}\right|^{2k}\right).$$

Proof. let $\{q_1, q_2, \cdots, q_n\}$ be the orthogonal basis of $\mathbb{R}$. Then v^0 can be rewritten as

$$v^0 = \sum \alpha_j q_j.$$

Moreover, following the power algorithm, we have

$$\begin{aligned}
w^1 &= Av^0 = \sum \alpha_j A q_j = \sum \alpha_j \lambda_j q_j . (Aq_j = \lambda_j q_j) \\
v^1 &= \frac{\sum \alpha_j \lambda_j q_j}{\sqrt{\sum \alpha_j^2 \lambda_j^2}} \\
w^2 &= Av^1 = \frac{\sum \alpha_j \lambda_j A q_j}{\sqrt{\sum \alpha_j^2 \lambda_j^2}} = \frac{\sum \alpha_j \lambda_j^2 q_j}{\sqrt{\sum \alpha_j^2 \lambda_j^2}} \\
v^2 &= \frac{\sum \alpha_j \lambda_j^2 q_j}{\sqrt{\sum \alpha_j^2 \lambda_j^{2\cdot 2}}} \\
&\cdots \\
w^k &= \frac{\sum \alpha_j \lambda_j^k q_j}{\sqrt{\sum \alpha_j^2 \lambda_j^{2\cdot(k-1)}}} \\
v^k &= \frac{\sum \alpha_j \lambda_j^k q_j}{\sqrt{\sum \alpha_j^2 \lambda_j^{2\cdot k}}}.
\end{aligned}$$

v^k can be rewritten as

$$\begin{aligned}
v^k &= \frac{\sum \alpha_j \lambda_j^k q_j}{\sqrt{\sum \alpha_j^2 \lambda_j^{2\cdot k}}} = \frac{\alpha_1 \lambda_1^k q_1 + \sum_{j>1} \alpha_j \lambda_j^k q_j}{\sqrt{\alpha_1^2 \lambda_1^{2k} + \sum_{j>1} \alpha_j^2 \lambda_j^{2k}}} \\
&= \frac{\alpha_1 \lambda_1^k}{|\alpha_1 \lambda_1^k|} \cdot \frac{q_1 + \sum_{j>1} \frac{\alpha_j}{\alpha_1} \left(\frac{\lambda_j}{\lambda_1}\right)^k q_j}{\sqrt{1 + \sum_{j>1} \left(\frac{\alpha_j}{\alpha_1}\right)^2 \left(\frac{\lambda_j}{\lambda_1}\right)^{2k}}} \\
&= \pm 1 \frac{q_1 + \sum_{j>1} \frac{\alpha_j}{\alpha_1} \left(\frac{\lambda_j}{\lambda_1}\right)^k q_j}{\sqrt{1 + \sum_{j>1} \left(\frac{\alpha_j}{\alpha_1}\right)^2 \left(\frac{\lambda_j}{\lambda_1}\right)^{2k}}}.
\end{aligned}$$

Therefore,

$$\left\| v^k - (\pm q_1) \right\|_2 \leq \left| \sum_{j>1} \frac{\alpha_j}{\alpha_1} \left(\frac{\lambda_j}{\lambda_1}\right)^k q_j \right| \leq C \left(\left| \frac{\lambda_2}{\lambda_1} \right| \right)^k = \mathcal{O}\left(\left| \frac{\lambda_2}{\lambda_1} \right|^k \right).$$

From Taylor formula

$$\left\| \lambda^k - \lambda_1 \right\|_2 = |\mathcal{R}(v^k) - \mathcal{R}(q_1)| = \mathcal{O} \left\| v^k - q_1 \right\|_2^2 = \mathcal{O}\left(\left| \frac{\lambda_2}{\lambda_1} \right|^{2k} \right).$$

□

Remark 4.3. *This shows that the speed of convergence depends on the gap between the two largest eigenvalues of A. In particular, if the largest eigenvalue of A were complex (which it can't be for the real symmetric matrices we are considering), then $\lambda_2 = \bar{\lambda}_1$ and the algorithm would not converge at all.*

4.4 Inverse Power iteration algorithm

Algorithm 4.5. *(inverse Power iteration algorithm)*

- v^0: *an arbitrary nonzero vector with* $\left\|v^0\right\|_2 = 1$
- *compute* $w^k = A^{-1}v^{k-1}$
- *compute* $v^k = \frac{w^k}{\|w^k\|_2}$
- *compute* $\lambda^k = \mathcal{R}(v^k)$

Algorithm 4.6. *(Improved inverse Power iteration algorithm)*

- v^0: *an arbitrary nonzero vector with* $\left\|v^0\right\|_2 = 1$
- *compute* $w^k = (A-\mu I)^{-1}v^{k-1}$
- *compute* $v^k = \frac{w^k}{\|w^k\|_2}$
- *compute* $\lambda^k = \mathcal{R}(v^k)$

Remark 4.4. *Improved inverse Power iteration algorithm is a shift μ.*

Algorithm 4.7. *(Rayleigh Quotient Iteration iteration algorithm)*

- v^0: *an arbitrary nonzero vector with* $\left\|v^0\right\|_2 = 1$
- *compute* $\lambda^0 = \mathcal{R}(v^0)$
- *compute* $w^k = (A-\lambda^{k-1}I)^{-1}v^{k-1}$
- *compute* $v^k = \frac{w^k}{\|w^k\|_2}$
- *compute* $\lambda^k = \mathcal{R}(v^k)$

Theorem 4.3. *(Convergence of power algorithm) If $A = A^*$, $q_1^T v^0 \neq 0$ and $|\lambda_1| > |\lambda_2| \geq \cdots |\lambda_m| \geq 0$, If we update the estimate μ for the eigenvalue with the Rayleigh quotient at each iteration we can get a cubically convergent algorithm, i.e.*

$$\left\|v^{k+1} - (\pm q_J)\right\|_2 = \mathcal{O}\left(\left\|v^k - (\pm q_J)\right\|_2^3\right)$$
$$\left\|\lambda^k - \lambda_J\right\|_2 = \mathcal{O}\left(|\lambda^k - \lambda_J|^3\right).$$

4.5 Problems

Problem 4.1. *(Prelim Aug. 2013#1)*

Solution. ◂

5 Solution of Nonlinear problems

Definition 5.1. *(convergence with Order p) An iterative scheme converges with order p>0 if there is a constant $C > 0$, such that*

$$|x - x^{k+1}| \leq C|x - x^k|^p. \tag{118}$$

5.1 Bisection method

Definition 5.2. *(Bisection method) The method is applicable for solving the equation f(x) = 0 for the real variable x, where f is a continuous function defined on an interval [a, b] and f(a) and f(b) have opposite signs i.e. $f(a)f(b) < 0$. In this case a and b are said to bracket a root since, by the intermediate value theorem, the continuous function f must have at least one root in the interval (a, b).*

Algorithm 1 Bisection method

$a_0 \leftarrow a, b_0 \leftarrow b$
while $k > 0$ **do**
  $c_k \leftarrow \frac{a_{k-1}+b_{k-1}}{2}$
  if $f(a_k)f(c_k) < 0$ **then**
    $a_k \leftarrow a_{k-1}$
    $b_k \leftarrow c_k$
  end if
  if $f(b_k)f(c_k) < 0$ **then**
    $a_k \leftarrow c_k$
    $b_k \leftarrow b_{k-1}$
  end if
  $x^k \leftarrow c^k \leftarrow \frac{a_k+b_k}{2}$
end while

5.2 Chord method

Definition 5.3. *(Chord method) The method is applicable for solving the equation f(x) = 0 for the real variable x, where f is a continuous function defined on an interval [a, b] and f(a) and f(b) have opposite signs i.e. $f(a)f(b) < 0$. Instead of the [a,b] segment halving, we?ll divide it relation $f(a) : f(b)$, It gives the approach of a root of the equation*

$$x^{k+1} = x^k - \left[\eta^k\right]^{-1} f(x^k).$$

where

$$\eta^k = \frac{f(b) - f(a)}{b - a}$$

Algorithm 2 Chord method

1: $x_1 = a - \frac{f(a)}{f(b)-f(a)}(b-a)$, $x_0 = 0$
2: $\eta^k = \frac{f(b)-f(a)}{b-a}$
3: **while** $|x^{k+1} - x^k| < \epsilon$ **do**
4: $\quad x^{k+1} \leftarrow x^k - \left[\eta^k\right]^{-1} f(x^k)$
5: **end while**

5.3 Secant method

Definition 5.4. *(Secant method) The method is applicable for solving the equation f(x) = 0 for the real variable x, where f is a continuous function defined on an interval [a, b] and f(a) and f(b) have opposite signs i.e. $f(a)f(b) < 0$. Instead of the [a,b] segment halving, we?ll divide it relation $f(x^k) : f(x^{k-1})$, It gives the approach of a root of the equation*

$$x^{k+1} = x^k - \left[\eta^k\right]^{-1} f(x^k).$$

where

$$\eta^k = \frac{f(x^k) - f(x^{k-1})}{x^k - x^{k-1}}$$

Algorithm 3 Secant method

1: $x_1 = a - \frac{f(a)}{f(b)-f(a)}(b-a)$
2: $\eta^k = \frac{f(x^k)-f(x^{k-1})}{x^k-x^{k-1}}, x_0 = 0$
3: **while** $|x^{k+1} - x^k| < \epsilon$ **do**
4: $\quad x^{k+1} \leftarrow x^k - \left[\eta^k\right]^{-1} f(x^k)$
5: **end while**

5.4 Newton's method

Definition 5.5. *(Newton's method) The method is applicable for solving the equation f(x) = 0 for the real variable x, where f is a continuous function defined on an interval [a, b] and f(a) and f(b) have opposite signs i.e. $f(a)f(b) < 0$. Instead of the [a,b] segment halving, we?ll divide it relation $f'(x^k)$, It gives the approach of a root of the equation*

$$x^{k+1} = x^k - \left[\eta^k\right]^{-1} f(x^k).$$

where

$$\eta^k = f'(x^k)$$

Remark 5.1. *This scheme needs $f'(x^k) \neq 0$.*

Algorithm 4 Newton's method

1: $x_1 = a - \frac{f(a)}{f(b)-f(a)}(b-a)$
2: $\eta^k = f'(x^k), x_0 = 0$
3: **while** $|x^{k+1} - x^k| < \epsilon$ **do**
4: $\quad x^{k+1} \leftarrow x^k - \left[\eta^k\right]^{-1} f(x^k)$
5: **end while**

Theorem 5.1. *(convergence of Newton's method) Let $f \in \mathbb{C}^2, f(x^*) = 0, f'(x) \neq 0$ and $f''(x^*)$ is bounded in a neighborhood of x^*. Provide x^0 is sufficient close to x^*, then newton's method converges quadratically, i.e.*

$$\left|x^{k+1} - x^*\right| \leq C\left|x^k - x^*\right|^2.$$

Proof. Let x^* be the root of $f(x)$. From the Taylor expansion, we know

$$0 = f(x^*) = f(x^k) + f'(x^k)(x^* - x^k) + \frac{1}{2}f''(\theta)(x^* - x^k)^2,$$

where θ is between x^* and x^k. Define $e^k = x^* - x^k$, then

$$0 = f(x^*) = f(x^k) + f'(x^k)(e^k) + \frac{1}{2}f''(\theta)(e^k)^2.$$

so

$$\left[f'(x^k)\right]^{-1} f(x^k) = -(e^k) - \frac{1}{2}\left[f'(x^k)\right]^{-1} f''(\theta)(e^k)^2.$$

From the Newton's scheme, we have

$$\begin{cases} x^{k+1} = x^k - \left[f'(x^k)\right]^{-1} f(x^k) \\ x^* = x^* \end{cases}$$

So,

$$e^{k+1} = e^k + \left[f'(x^k)\right]^{-1} f(x^k) = -\frac{1}{2}\left[f'(x^k)\right]^{-1} f''(\theta)(e^k)^2,$$

i.e.

$$e^{k+1} = -\frac{f''(\theta)}{2\left[f'(x^k)\right]}(e^k)^2,$$

By assumption, there is a neighborhood of x, such that

$$\left|f(z)\right| \leq C_1, \quad \left|f'(z)\right| \leq C_2,$$

Therefore,

$$\left|e^{k+1}\right| \leq \frac{\left|f''(\theta)\right|}{\left|2\left[f'(x^k)\right]\right|}(e^k)^2 \leq \frac{C_1}{2C_2}\left|e^k\right|^2.$$

This implies

$$\left|x^{k+1} - x^*\right| \leq C\left|x^k - x^*\right|^2.$$

□

5.5 Newton's method for system

Theorem 5.2. *If $F:\mathbb{R}\to\mathbb{R}^n$ is integrable over the interval [a,b], then*

$$\left\|\int_a^b F(t)\,dt\right\| \le \int_a^b \|F(t)\|\,dt.$$

Theorem 5.3. *Suppose $F:\mathbb{R}^n\to\mathbb{R}^m$ is continuously differentiable and $a,b\in\mathrm{R}^n$. Then*

$$F(b)=F(a)+\int_0^1 J(a+\theta(b-a))(b-a)\,d\theta,$$

where J is the Jacobian of F.

Theorem 5.4. *Suppose $J:\mathbb{R}^m\to\mathbb{R}^{n\times n}$ is a continuous matrix-valued function. If $J(x^*)$ is nonsingular, then there exists $\delta>0$ such that, for all $x\in\mathrm{R}^m$ with $\|x-x^*\|<\delta$, $J(x)$ is nonsingular and*

$$\left\|J(x)^{-1}\right\| < 2\left\|J(x^*)^{-1}\right\|.$$

Theorem 5.5. *Suppose $J:\mathbb{R}^n\to\mathbb{R}^m$. Then F is said to be Lipschitz continuous on $S\subset\mathbb{R}^n$ if there exists a positive constant L such that*

$$\left\|J(x)-J(y)\right\| \le L\left\|x-y\right\|$$

Theorem 5.6. *(convergence of Newton's method) Suppose $F:\mathbb{R}^n\to\mathbb{R}^n$ is continuously differentiable and $F(\mathbf{x}^*)=\mathbf{0}$.*

1. *the Jacobian $J(\mathbf{x}^*)$ of F at $\mathbf{x}^*$ is nonsingular, and*
2. *J is Lipschitz continuous on a neighborhood of $\mathbf{x}^*$,*

then, for all $\mathbf{x^0}$ sufficiently close to $\mathbf{x}^*$, $\left|\mathbf{x^0}-\mathbf{x}^*\right|<\epsilon$, Newton's method converges quadratically to $\mathbf{x}^*$, i.e

$$\left|\mathbf{x^{k+1}}-\mathbf{x^k}\right| \le C\left|\mathbf{x^k}-\mathbf{x}\right|^2.$$

Proof. Let $\mathbf{x}^*$ be the root of $F(x)$ i.e. $F(\mathbf{x}^*)$=0. From the Newton's scheme, we have

$$\begin{cases}\mathbf{x^{k+1}}=\mathbf{x^k}-\left[\mathbf{J}(\mathbf{x^k})\right]^{-1}\mathbf{F}(\mathbf{x^k})\\ \mathbf{x}^*=\mathbf{x}^*\end{cases}$$

Therefore, we have

$$\begin{aligned}\mathbf{x}^*-\mathbf{x^{k+1}} &= \mathbf{x}^*-\mathbf{x^k}+\left[\mathbf{J}(\mathbf{x^k})\right]^{-1}(\mathbf{F}(\mathbf{x^k})-\mathbf{F}(\mathbf{x}^*))\\ &= \mathbf{x}^*-\mathbf{x^k}+\left[\mathbf{J}(\mathbf{x^k})\right]^{-1}\left(\mathbf{F}(\mathbf{x^k})-\mathbf{F}(\mathbf{x}^*)+\mathbf{J}(\mathbf{x}^*)(\mathbf{x}^*-\mathbf{x^k})-\mathbf{J}(\mathbf{x}^*)(\mathbf{x}^*-\mathbf{x^k})\right)\\ &= \left(I-\left[J(\mathbf{x^k})\right]^{-1}J(\mathbf{x}^*)\right)\left(\mathbf{x}^*-\mathbf{x^k}\right)-\left[J(\mathbf{x^k})\right]^{-1}\left(F(\mathbf{x^k})-\mathbf{F}(\mathbf{x}^*)+\mathbf{J}(\mathbf{x}^*)(\mathbf{x}^*-\mathbf{x^k})\right).\end{aligned}$$

So,

$$\left\|\mathbf{x}^*-\mathbf{x^{k+1}}\right\| \le \left\|I-\left[J(\mathbf{x^k})\right]^{-1}J(\mathbf{x}^*)\right\|\left\|\mathbf{x}^*-\mathbf{x^k}\right\|+\left\|\left[J(\mathbf{x^k})\right]^{-1}\right\|\left\|F(\mathbf{x^k})-\mathbf{F}(\mathbf{x}^*)+\mathbf{J}(\mathbf{x}^*)(\mathbf{x}^*-\mathbf{x^k})\right\|. \tag{119}$$

Now, we will estimate $\left\|I-\left[J(\mathbf{x^k})\right]^{-1}J(\mathbf{x}^*)\right\|$ and $\left\|F(\mathbf{x^k})-\mathbf{F}(\mathbf{x}^*)+\mathbf{J}(\mathbf{x}^*)(\mathbf{x}^*-\mathbf{x^k})\right\|$.

$$\begin{aligned}\left\|I-\left[J(\mathbf{x^k})\right]^{-1}J(\mathbf{x}^*)\right\| &= \left\|\left[J(\mathbf{x^k})\right]^{-1}\left[J(\mathbf{x^k})\right]-\left[J(\mathbf{x^k})\right]^{-1}J(\mathbf{x}^*)\right\| \\ &= \left\|\left[J(\mathbf{x^k})\right]^{-1}\left(J(\mathbf{x^k})-\mathbf{J}(\mathbf{x}^*)\right)\right\| \\ &\leq \left\|\left[J(\mathbf{x^k})\right]^{-1}\right\|\left\|J(\mathbf{x^k})-\mathbf{J}(\mathbf{x}^*)\right\| \\ &\leq L\left\|\left[J(\mathbf{x^k})\right]^{-1}\right\|\left\|\mathbf{x}^*-\mathbf{x^k}\right\|.\end{aligned} \tag{120}$$

In the last step of the above equation, we use the J is Lipschitz continuous(If J is not Lipschitz continuous, we can only get the Newton method converges linearly to $\mathbf{x}^*$). Since $F:\mathbb{R}^n\to\mathbb{R}^n$ is continuously differentiable, therefore

$$F(b)=F(a)+\int_0^1 J(a+\theta(b-a))(b-a)d\theta.$$

So

$$\begin{aligned}F(\mathbf{x^k}) &= F(\mathbf{x}^*)+\int_{\mathbf{0}}^{\mathbf{1}}\mathbf{J}(\mathbf{x}^*+\theta(\mathbf{x^k}-\mathbf{x}^*))(\mathbf{x^k}-\mathbf{x}^*)\mathbf{d}\theta \\ &= F(\mathbf{x}^*)+\int_{\mathbf{0}}^{\mathbf{1}}\mathbf{J}(\mathbf{x}^*+\theta(\mathbf{x^k}-\mathbf{x}^*))(\mathbf{x^k}-\mathbf{x}^*)+\mathbf{J}(\mathbf{x}^*)(\mathbf{x}^*-\mathbf{x^k})-\mathbf{J}(\mathbf{x}^*)(\mathbf{x}^*-\mathbf{x^k})\mathbf{d}\theta \\ &= F(\mathbf{x}^*)-\mathbf{J}(\mathbf{x}^*)(\mathbf{x}^*-\mathbf{x^k})+\int_{\mathbf{0}}^{\mathbf{1}}\mathbf{J}(\mathbf{x}^*+\theta(\mathbf{x^k}-\mathbf{x}^*))(\mathbf{x^k}-\mathbf{x}^*)+\mathbf{J}(\mathbf{x}^*)(\mathbf{x}^*-\mathbf{x^k})\mathbf{d}\theta\end{aligned}$$

Hence

$$F(\mathbf{x^k})-\mathbf{F}(\mathbf{x}^*)+\mathbf{J}(\mathbf{x}^*)(\mathbf{x}^*-\mathbf{x^k})=\int_{\mathbf{0}}^{\mathbf{1}}\mathbf{J}(\mathbf{x}^*+\theta(\mathbf{x^k}-\mathbf{x}^*))(\mathbf{x^k}-\mathbf{x}^*)+\mathbf{J}(\mathbf{x}^*)(\mathbf{x}^*-\mathbf{x^k})\mathbf{d}\theta.$$

So,

$$\begin{aligned}\left\|F(\mathbf{x^k})-\mathbf{F}(\mathbf{x}^*)+\mathbf{J}(\mathbf{x}^*)(\mathbf{x}^*-\mathbf{x^k})\right\| &= \left\|\int_0^1 J(\mathbf{x}^*+\theta(\mathbf{x^k}-\mathbf{x}^*))(\mathbf{x^k}-\mathbf{x}^*)+\mathbf{J}(\mathbf{x}^*)(\mathbf{x}^*-\mathbf{x^k})\mathbf{d}\theta\right\| \\ &\leq \int_0^1\left\|J(\mathbf{x}^*+\theta(\mathbf{x^k}-\mathbf{x}^*))(\mathbf{x^k}-\mathbf{x}^*)+\mathbf{J}(\mathbf{x}^*)(\mathbf{x}^*-\mathbf{x^k})\right\|d\theta \\ &\leq \int_0^1\left\|J(\mathbf{x}^*+\theta(\mathbf{x^k}-\mathbf{x}^*))-\mathbf{J}(\mathbf{x}^*)\right\|\left\|(\mathbf{x}^*-\mathbf{x^k})\right\|d\theta \\ &\leq \int_0^1 L\theta\left\|(\mathbf{x}^*-\mathbf{x^k})\right\|^2 d\theta \\ &\leq \frac{1}{2}L\left\|(\mathbf{x}^*-\mathbf{x^k})\right\|^2.\end{aligned} \tag{121}$$

From (119), (120) and (121), we have

$$\left\|\mathbf{x}^*-\mathbf{x^{k+1}}\right\|\leq\frac{3}{2}L\left\|\left[J(\mathbf{x^k})\right]^{-1}\right\|\left\|\mathbf{x}^*-\mathbf{x^k}\right\|^2\leq 3L\left\|[J(\mathbf{x}^*)]^{-1}\right\|\left\|\mathbf{x}^*-\mathbf{x^k}\right\|^2. \tag{122}$$

□

Remark 5.2. *From the last step of the above proof process, we can get the condition of ϵ. such as, If*

$$\left\|\mathbf{x}^* - \mathbf{x}^{\mathbf{k}}\right\| \leq \frac{1}{L\left\|[J(\mathbf{x}^*)]^{-1}\right\|},$$

then

$$\left\|\mathbf{x}^* - \mathbf{x}^{\mathbf{k+1}}\right\| \leq \frac{1}{2}\left\|\mathbf{x}^* - \mathbf{x}^{\mathbf{k}}\right\|. \tag{123}$$

5.6 Fixed point method

In fact, Chord, scant and Newton's method can be consider as fixed point iterative, since

$$x^{k+1} = x^k - \left[\eta^k\right]^{-1} f(x^k) = \phi(x^k).$$

Theorem 5.7. *x is a fixed point of ϕ and $U_\delta = \{z : |x - z| \leq \delta\}$. If ϕ is differentiable on U_δ and $q < 1$ such $\left|\phi'(z)\right| \leq q < 1$ for all $z \in U_\delta$, then*

1. *$\phi(U_\delta) \subset U_\delta$*
2. *ϕ is contraction.*

5.7 Problems

Problem 5.1. *(Prelim Jan. 2011#4) Let $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}^n$ be twice continuously differentiable. Suppose $x^* \in \Omega$ is a solution of $f(x) = 0$, and the Jacobian matrix of f, denoted J_f, is invertible at x^*.*

1. *Prove that if $x^0 \in \Omega$ is sufficiently close to x^*, then the following iteration converges to x^*:*

$$x^{k+1} = x^k - J_f(x^0)^{-1} f(x^k).$$

2. *Prove that the convergence is typically only linear.*

Solution. Let $\mathbf{x}^*$ be the root of $\mathbf{f}(x)$ i.e. $\mathbf{f}(\mathbf{x}^*)$=0. From the Newton's scheme, we have

$$\begin{cases} \mathbf{x}^{k+1} = \mathbf{x}^k - \left[J(\mathbf{x}^0)\right]^{-1} \mathbf{f}(\mathbf{x}^k) \\ \mathbf{x}^* = \mathbf{x}^* \end{cases}$$

Therefore, we have

$$\begin{aligned} \mathbf{x}^* - \mathbf{x}^{k+1} &= \mathbf{x}^* - \mathbf{x}^k + \left[J(\mathbf{x}^0)\right]^{-1} (\mathbf{f}(\mathbf{x}^k) - \mathbf{f}(\mathbf{x}^*)) \\ &= \mathbf{x}^* - \mathbf{x}^k - \left[\mathbf{J}(\mathbf{x}^0)\right]^{-1} \mathbf{J}(\xi)(\mathbf{x}^* - \mathbf{x}^k). \end{aligned}$$

therefore

$$\left|\mathbf{x}^* - \mathbf{x}^{k+1}\right| \leq \left|1 - \frac{\mathbf{J}(\xi)}{\mathbf{J}(\mathbf{x}^0)}\right| \left|\mathbf{x}^* - \mathbf{x}^k\right|$$

From theorem

Theorem 5.8. *Suppose $J : \mathbb{R}^m \to \mathbb{R}^{n\times n}$ is a continuous matrix-valued function. If $J(x^*)$ is nonsingular, then there exists $\delta > 0$ such that, for all $x \in \mathrm{R}^m$ with $\|x - x^*\| < \delta$, $J(x)$ is nonsingular and*

$$\left\|J(x)^{-1}\right\| < 2\left\|J(x^*)^{-1}\right\|.$$

we get

$$\left|\mathbf{x}^* - \mathbf{x}^{k+1}\right| \leq \frac{1}{2}\left|\mathbf{x}^* - \mathbf{x}^k\right|.$$

Which also shows the convergence is typically only linear.

◀

Problem 5.2. *(Prelim Aug. 2010#5) Assume that $f : \mathbb{R} \to \mathbb{R}, f \in C^2(\mathbb{R}), f'(x) > 0$ for all $x \in \mathbb{R}$, and $f''(x) > 0$, for all $x \in \mathbb{R}$.*

1. *Suppose that a root $\xi \in \mathbb{R}$ exists. Prove that it is unique. Exhibit a function satisfying the assumptions above that has no root.*
2. *Prove that for any starting guess $x_0 \in \mathbb{R}$, Newton's method converges, and the convergence rate is quadratic.*

Solution. 1. Let x_1 and x_2 are the two different roots. So, $f(x_1) = f(x_2) = 0$, then by Mean value theorem, we have that there exists $\eta \in [x_1, x_2]$, such $f'(\eta) = 0$ which contradicts $f'(x) > 0$.

2. example $f(x) = e^x$.

3. Let x^* be the root of $f(x)$. From the Taylor expansion, we know

$$0 = f(x^*) = f(x^k) + f'(x^k)(x^* - x^k) + \frac{1}{2}f''(\theta)(x^* - x^k)^2,$$

where θ is between x^* and x^k. Define $e^k = x^* - x^k$, then

$$0 = f(x^*) = f(x^k) + f'(x^k)(e^k) + \frac{1}{2}f''(\theta)(e^k)^2.$$

so

$$\left[f'(x^k)\right]^{-1} f(x^k) = -(e^k) - \frac{1}{2}\left[f'(x^k)\right]^{-1} f''(\theta)(e^k)^2.$$

From the Newton's scheme, we have

$$\begin{cases} x^{k+1} = x^k - \left[f'(x^k)\right]^{-1} f(x^k) \\ x^* = x^* \end{cases}$$

So,

$$e^{k+1} = e^k + \left[f'(x^k)\right]^{-1} f(x^k) = -\frac{1}{2}\left[f'(x^k)\right]^{-1} f''(\theta)(e^k)^2,$$

i.e.

$$e^{k+1} = -\frac{f''(\theta)}{2\left[f'(x^k)\right]}(e^k)^2,$$

By assumption, there is a neighborhood of x, such that

$$\left|f(z)\right| \leq C_1, \quad \left|f'(z)\right| \leq C_2,$$

Therefore,

$$\left|e^{k+1}\right| \leq \frac{\left|f''(\theta)\right|}{\left|2\left[f'(x^k)\right]\right|}(e^k)^2 \leq \frac{C_1}{2C_2}\left|e^k\right|^2.$$

This implies

$$\left|x^{k+1}-x^*\right| \leq C\left|x^k-x^*\right|^2.$$

◀

Problem 5.3. *(Prelim Aug. 2010#4) Let $f:\mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable. Suppose x^* is a isolated root of f and the Jacobian of f at $x^*(J(x^*))$ is non-singular. Determine conditions on ϵ so that if $\|x_0-x^*\|_2 < \epsilon$ then the following iteration converges to x^*:*

$$x_{k+1} = x_k - J_f(x_0)^{-1}f(x_k), \quad k=0,1,2,\cdots.$$

Solution. ◀

Problem 5.4. *(Prelim Aug. 2009#5) Consider the two-step Newton method*

$$y_k = x_k - \frac{f(x_k)}{f'(x_k)}, x_{k+1} = y_k - \frac{f(y_k)}{f'(x_k)}$$

for the solution of the equation $f(x)=0$. Prove

1. *If the method converges, then*

$$\lim_{k\to\infty}\frac{x_{k+1}-x^*}{(y_k-x^*)(x_k-x^*)} = \frac{f''(x_k)}{f'(x_k)},$$

   where x^* is the solution.

2. *Prove the convergence is cubic, that is*

$$\lim_{k\to\infty}\frac{x_{k+1}-x^*}{(x_k-x^*)^3} = \frac{1}{2}\left(\frac{f''(x_k)}{f'(x_k)}\right).$$

3. *Would you say that this method is faster than Newton's method given that its convergence is cubic?*

Solution. 1. First, we will show that if $x_k \in [x-h, x+h]$, then $y_k \in [x-h, x+h]$. By Taylor expansion formula, we have

$$0 = f(x^*) = f(x_k) + f'(x_k)(x^*-x_k) + \frac{1}{2!}f''(\xi_k)(x^*-x_k)^2,$$

where ξ is between x and x_k. Therefore, we have

$$f(x_k) = -f'(x_k)(x^*-x_k) - \frac{1}{2!}f''(\xi_k)(x^*-x_k)^2.$$

Plugging the above equation to the first step of the Newton's method, we have

$$y_k = x_k + (x^*-x_k) + \frac{1}{2!}\frac{f''(\xi_k)}{f'(x_k)}(x^*-x_k)^2.$$

then

$$y_k - x^* = \frac{1}{2!}\frac{f''(\xi_k)}{f'(x_k)}(x^* - x_k)^2. \tag{124}$$

Therefore,

$$\left|y_k - x^*\right| = \left|\frac{1}{2!}\frac{f''(\xi_k)}{f'(x_k)}(x^* - x_k)^2\right| \leq \frac{1}{2}\left|\frac{f''(\xi_k)}{f'(x_k)}\right|\left|(x^* - x_k)\right|\left|(x^* - x_k)\right|.$$

Since we can choose the initial value very close to x^*, shah that

$$\left|\frac{f''(\xi)}{f'(x_k)}\right|\left|(x^* - x_k)\right| \leq 1$$

Then, we have that

$$\left|y_k - x^*\right| \leq \frac{1}{2}\left|(x^* - x_k)\right|.$$

Hence, we proved the result, that is to say, if $x_k \to x^*$, then $y_k, \xi_k \to x^*$.

2. Next, we will show if $x_k \in [x-h, x+h]$, then $x_{k+1} \in [x-h, x+h]$. From the second step of the Newton's Method, we have that

$$\begin{aligned} x_{k+1} - x^* &= y_k - x^* - \frac{f(y_k)}{f'(x_k)} \\ &= \frac{1}{f'(x_k)}\left((y_k - x^*)f'(x_k) - f(y_k)\right) \\ &= \frac{1}{f'(x_k)}\left[(y_k - x^*)(f'(x_k) - f'(x^*)) - f(y_k) + (y_k - x^*)f'(x^*)\right] \end{aligned}$$

By mean value theorem, we have there exists η_k between x^* and x_k, such that

$$f'(x_k) - f'(x^*) = f''\eta_k(x_k - x^*),$$

and by Taylor expansion formula, we have

$$\begin{aligned} f(y_k) &= f(x^*) + f'(x^*)(y_k - x^*) + \frac{(y_k - x^*)^2}{2}f''(\gamma_k) \\ &= f'(x^*)(y_k - x^*) + \frac{(y_k - x^*)^2}{2}f''(\gamma_k), \end{aligned}$$

where γ is between y_k and x^*. Plugging the above two equations to the second step of the Newton's method, we get

$$\begin{aligned} x_{k+1} - x^* &= \frac{1}{f'(x_k)}\left[f''\eta_k(x_k - x^*)(y_k - x^*) - f'(x^*)(y_k - x^*) - \frac{(y_k - x^*)^2}{2}f''(\gamma_k) + (y_k - x^*)f'(x^*)\right] \\ &= \frac{1}{f'(x_k)}\left[f''\eta_k(x_k - x^*)(y_k - x^*) - \frac{(y_k - x^*)^2}{2}f''(\gamma_k)\right]. \end{aligned} \tag{125}$$

Taking absolute values of the above equation, then we have

$$\begin{aligned} \left|x_{k+1} - x^*\right| &= \left|\frac{1}{f'(x_k)}\left[f''\eta_k(x_k - x^*)(y_k - x^*) - \frac{(y_k - x^*)^2}{2}f''(\gamma_k)\right]\right| \\ &\leq A\left|x_k - x^*\right|\left|y_k - x^*\right| + \frac{A}{2}\left|y_k - x^*\right|\left|y_k - x^*\right| \\ &\leq \frac{1}{2}\left|x_k - x^*\right| + \frac{1}{8}\left|x_k - x^*\right| = \frac{5}{8}\left|x_k - x^*\right|. \end{aligned}$$

Hence, we proved the result, that is to say, if $y_k \to x^*$, then $x_{k+1}, \eta_k, \gamma_k \to x^*$.

3. Finally, we will prove the convergence order is cubic. From (215), we can get that

$$\frac{x_{k+1}-x^*}{(x_k-x^*)(y_k-x^*)}=\frac{f''\eta_k}{f'(x_k)}-\frac{(y_k-x^*)f''(\gamma_k)}{2(x_k-x^*)f'(x_k)}.$$

By using (214), we have

$$\frac{x_{k+1}-x^*}{(x_k-x^*)(y_k-x^*)}=\frac{f''\eta_k}{f'(x_k)}-\frac{1}{4}\frac{f''(\xi_k)}{f'(x_k)}(x^*-x_k)\frac{f''(\gamma_k)}{f'(x_k)}.$$

Taking limits gives

$$\lim_{k\to\infty}\frac{x_{k+1}-x^*}{(x_k-x^*)(y_k-x^*)}=\frac{f''(x^*)}{f'(x^*)}.$$

By using (214) again, we have

$$\frac{1}{y_k-x^*}=\frac{2}{(x^*-x_k)^2}\frac{f'(x_k)}{f''(\xi_k)}.$$

Hence

$$\lim_{k\to\infty}\frac{x_{k+1}-x^*}{(x_k-x^*)^3}=\frac{1}{2}\left(\frac{f''(x^*)}{f'(x^*)}\right)^2.$$

◀

6 Euler Method

In this section, we focus on

$$\begin{cases} y' = f(t,y), \\ y(t_0) = y_0. \end{cases}$$

Where f is Lipschitz continuous w.r.t. the second variable, i.e

$$|f(t,x) - f(t,y)| \le \lambda |x-y|, \quad \lambda > 0. \tag{126}$$

In the following, We will let $y(t_n)$ to be the numerical approximation of y_n and $e_n = y_n - y(t_n)$ to be the error.

Definition 6.1. *(Order of the Method) A time stepping scheme*

$$y_{n+1} = \Phi(h, y_0, y_1, \cdots, y_n) \tag{127}$$

is of order of $p \ge 1$, if

$$y_{n+1} - \Phi(h, y_0, y_1, \cdots, y_n) = \mathcal{O}(h^{p+1}). \tag{128}$$

Definition 6.2. *(Convergence of the Method) A time stepping scheme*

$$y_{n+1} = \Phi(h, y_0, y_1, \cdots, y_n) \tag{129}$$

is convergent , if

$$\lim_{h \to 0} \max_n \left\| y(t_n) - y_n \right\| = 0. \tag{130}$$

6.1 Euler's method

Definition 6.3. *(Forward Euler Method[a])*

$$y_{n+1} = y_n + h f(t_n, y_n), \quad n = 0, 1, 2, \cdots. \tag{131}$$

[a] Forward Euler Method is explicit.

Theorem 6.1. *(Forward Euler Method is of order 1 [a]) Forward Euler Method*

$$y(t_{n+1}) = y(t_n) + h f(t_n, y(t_n)), \tag{132}$$

is of order 1 .

[a] You can also use multi-step theorem to derive it.

Proof. By the Taylor expansion,

$$y(t_{n+1}) = y(t_n) + h y'(t_n) + \mathcal{O}(h^2). \tag{133}$$

So,

$$\begin{aligned} y(t_{n+1}) - y(t_n) - h f(t_n, y(t_n)) &= y(t_n) + h y'(t_n) + \mathcal{O}(h^2) - y(t_n) - h f(t_n, y(t_n)) \\ &= y(t_n) + h y'(t_n) + \mathcal{O}(h^2) - y(t_n) - h y'(t_n) \\ &= \mathcal{O}(h^2). \end{aligned} \tag{134}$$

Therefore, Forward Euler Method (6.3) is order of 1 . □

Theorem 6.2. *(The convergence of Forward Euler Method) Forward Euler Method*

$$y(t_{n+1}) = y(t_n) + hf(t_n, y(t_n)), \tag{135}$$

is convergent.

Proof. From (134), we get

$$y(t_{n+1}) = y(t_n) + hf(t_n, y(t_n)) + \mathcal{O}(h^2), \tag{136}$$

Subtracting (136) from (131), we get

$$e_{n+1} = e_n + h[f(t_n, y_n) - f(t_n, y(t_n))] + ch^2. \tag{137}$$

Since f is lipschitz continuous w.r.t. the second variable, then

$$|f(t_n, y_n) - f(t_n, y(t_n))| \le \lambda |y_n - y(t_n)|, \quad \lambda > 0. \tag{138}$$

Therefore,

$$\begin{aligned} \|e_{n+1}\| &\le \|e_n\| + h\lambda \|e_n\| + ch^2 \\ &= (1 + h\lambda)\|e_n\| + ch^2. \end{aligned} \tag{139}$$

Claim:[2]

$$\|e_n\| \le \frac{c}{\lambda} h[(1 + h\lambda)^n - 1], n = 0, 1, \cdots \tag{140}$$

Proof for Claim (221): The proof is by induction on n.

1. when $n = 0$, $e_n = 0$, hence $\|e_n\| \le \frac{c}{\lambda} h[(1 + h\lambda)^n - 1]$,
2. Induction assumption:

$$\|e_n\| \le \frac{c}{\lambda} h[(1 + h\lambda)^n - 1]$$

3. Induction steps:

$$\begin{aligned} \|e_{n+1}\| &\le (1 + h\lambda)\|e_n\| + ch^2 && (141) \\ &\le (1 + h\lambda)\frac{c}{\lambda} h[(1 + h\lambda)^n - 1] + ch^2 && (142) \\ &= \frac{c}{\lambda} h[(1 + h\lambda)^{n+1} - 1]. && (143) \end{aligned}$$

So, from the claim (221), we get $\|e_n\| \to 0$, when $h \to 0$. Therefore Forward Euler Method is convergent . □

Definition 6.4. *(tableaux) The tableaux of Forward Euler method*

0	0
	1.

Solution. Since, the Forward Euler method is as follows

$$y_{n+1} = y_n + hf(t_n, y_n),$$

then it can be rewritten as RK format, i.e.

$$\begin{aligned} \xi_1 &= y_n \\ y_{n+1} &= y_n + hf(t_n + 0h, \xi_1). \end{aligned}$$

◀

Definition 6.5. *(Backward Euler Methods[a])*

$$y_{n+1} = y_n + hf(t_{n+1}, y_{n+1}), \quad n = 0, 1, 2, \cdots. \tag{144}$$

[a] Backward Euler Method is implicit.

Theorem 6.3. *(backward Euler Method is of order 1 [a])* *Backward Euler Method*

$$y(t_{n+1}) = y(t_n) + hf(t_{n+1}, y(t_{n+1})), \tag{145}$$

is of order 1 .

[a] You can also use multi-step theorem to derive it.

Proof. By the Taylor expansion,

$$\begin{aligned} y(t_{n+1}) &= y(t_n) + hy'(t_n) + \mathcal{O}(h^2) && (146) \\ y'(t_{n+1}) &= y'(t_n) + \mathcal{O}(h). && (147) \end{aligned}$$

So,

$$\begin{aligned} & y(t_{n+1}) - y(t_n) - hf(t_{n+1}, y(t_{n+1})) \\ = \; & y(t_{n+1}) - y(t_n) + hy'(t_{n+1}) \\ = \; & y(t_n) + hy'(t_n) + \mathcal{O}(h^2) - y(t_n) - h[y'(t_n) + \mathcal{O}(h)] && (148) \\ = \; & \mathcal{O}(h^2). \end{aligned}$$

Therefore, Backward Euler Method (6.5) is order of 1 . □

Theorem 6.4. *(The convergence of Backward Euler Method)* *Backward Euler Method*

$$y(t_{n+1}) = y(t_n) + hf(t_{n+1}, y(t_{n+1})), \tag{149}$$

is convergent.

Proof. From (148), we get

$$y(t_{n+1}) = y(t_n) + hf(t_{n+1}, y(t_{n+1})) + \mathcal{O}(h^2), \tag{150}$$

Subtracting (150) from (144), we get

$$e_{n+1} = e_n + h[f(t_{n+1}, y_{n+1}) - f(t_{n+1}, y(t_{n+1}))] + ch^2. \tag{151}$$

Since f is lipschitz continuous w.r.t. the second variable, then

$$|f(t_{n+1},y_{n+1})-f(t_{n+1},y(t_{n+1}))|\leq\lambda|y_{n+1}-y(t_{n+1})|,\quad\lambda>0. \tag{152}$$

Therefore,

$$\left\|e_{n+1}\right\| \leq \|e_n\|+h\lambda\left\|e_{n+1}\right\|+ch^2. \tag{153}$$

So,

$$(1-h\lambda)\left\|e_{n+1}\right\| \leq \|e_n\|+ch^2. \tag{154}$$

So, by the Discrete Gronwall's Inequality , we have

$$\begin{aligned}\left\|e_{n+1}\right\| &\leq \frac{\|e_0\|}{(1-h\lambda)^{n+1}}+c\sum_{k=0}^{n}\frac{h^2}{(1-h\lambda)^{n+k-1}}\\ &= c\sum_{k=0}^{n}\frac{h^2}{(1-h\lambda)^{n+k-1}}\\ &\leq ch^2(1+h\lambda)^{(nh)/h\lambda}(1-h\lambda\to 1+h\lambda)\\ &\leq che^T T.\end{aligned} \tag{155}$$

So, from the claim (155), we get $\|e_n\|\to 0$, when $h\to 0$. Therefore Backward Euler Method is convergent . □

Definition 6.6. *(tableaux) The tableaux of Backward Euler method*

0	0	0
1	0	1
	0	1.

Solution. Since, the Backward Euler method is as follows

$$y_{n+1}=y_n+hf(t_{n+1},y_{n+1}),$$

then it can be rewritten as RK format, i.e.

$$\begin{aligned}\xi_1 &= y_n\\ \xi_2 &= y_n+h\left[0f(t_n+0h,\xi_1)+1f(t_n+1h,\xi_2)\right]\\ y_{n+1} &= y_n+hf(t_n+h,\xi_2).\end{aligned}$$

◀

6.2 Trapezoidal Method

Definition 6.7. *(Trapezoidal Method[a])*

$$y_{n+1}=y_n+\frac{1}{2}h[f(t_n,y_n)+f(t_{n+1},y_{n+1})],\quad n=0,1,2,\cdots. \tag{156}$$

[a] Trapezoidal Method Method is a combination of Forward Euler Method and Backward Euler Method.

Theorem 6.5. *(Trapezoidal Method is of order 2 [a])* *Trapezoidal Method*

$$y(t_{n+1}) = y(t_n) + \frac{1}{2}h[f(t_n, y(t_n)) + f(t_{n+1}, y(t_{n+1}))], \tag{157}$$

is of order 2 .

[a] You can also use multi-step theorem to derive it.

Proof. By the Taylor expansion,

$$\begin{aligned} y(t_{n+1}) &= y(t_n) + hy'(t_n) + \frac{1}{2!}h^2y''(t_n) + \mathcal{O}(h^3) && (158)\\ y'(t_{n+1}) &= y'(t_n) + hy''(t_n) + \mathcal{O}(h^2). && (159) \end{aligned}$$

So,

$$\begin{aligned} & y(t_{n+1}) - y(t_n) + \frac{1}{2}h[f(t_n, y(t_n)) + f(t_{n+1}, y(t_{n+1}))] \\ = \;& y(t_{n+1}) - y(t_n) + \frac{1}{2}h[y'(t_n) + y'(t_{n+1})] \\ = \;& y(t_n) + hy'(t_n) + \frac{1}{2!}h^2y''(t_n) + \mathcal{O}(h^3) - y(t_n) + \frac{1}{2}h[y'(t_n) + y'(t_n) + hy''(t_n) + \mathcal{O}(h^2)] \quad (160)\\ = \;& \mathcal{O}(h^3). \end{aligned}$$

Therefore, Trapezoidal Method (6.7) is order of 2 . □

Theorem 6.6. *(The convergence of Trapezoidal Method)* *Trapezoidal Method*

$$y(t_{n+1}) = y(t_n) + \frac{1}{2}h[f(t_n, y(t_n)) + f(t_{n+1}, y(t_{n+1}))], \tag{161}$$

is convergent.

Proof. From (160), we get

$$y(t_{n+1}) = y(t_n) + \frac{1}{2}h[f(t_n, y(t_n)) + f(t_{n+1}, y(t_{n+1}))] + \mathcal{O}(h^3), \tag{162}$$

Subtracting (162) from (156), we get

$$e_{n+1} = e_n + \frac{1}{2}h[f(t_n, y_n) - f(t_n, y(t_n)) + f(t_{n+1}, y_{n+1}) - f(t_{n+1}, y(t_{n+1}))] + ch^3. \tag{163}$$

Since f is lipschitz continuous w.r.t. the second variable, then

$$\begin{aligned} |f(t_n, y_n) - f(t_n, y(t_n))| &\le \lambda|y_n - y(t_n)|, \quad \lambda > 0, && (164)\\ |f(t_{n+1}, y_{n+1}) - f(t_{n+1}, y(t_{n+1}))| &\le \lambda|y_{n+1} - y(t_{n+1})|, \quad \lambda > 0. && (165) \end{aligned}$$

Therefore,

$$\left\|e_{n+1}\right\| \le \|e_n\| + \frac{1}{2}h\lambda(\|e_n\| + \left\|e_{n+1}\right\|) + ch^3. \tag{166}$$

So,

$$(1-\frac{1}{2}h\lambda)\left\|e_{n+1}\right\| \quad \leq \quad (1+\frac{1}{2}h\lambda)\|e_n\|+ch^3. \tag{167}$$

Claim:[2]

$$\|e_n\| \leq \frac{c}{\lambda}h^2\left[\left(\frac{1+\frac{1}{2}h\lambda}{1-\frac{1}{2}h\lambda}\right)^n-1\right], n=0,1,\cdots \tag{168}$$

Proof for Claim (168): The proof is by induction on n.

Then, we can make h small enough to such that $0<h\lambda<2$, then

$$\frac{1+\frac{1}{2}h\lambda}{1-\frac{1}{2}h\lambda}=1+\frac{1}{1-\frac{1}{2}h\lambda}\leq\sum_{\ell=0}^{\infty}\frac{1}{\ell!}\left(\frac{h\lambda}{1-\frac{1}{2}h\lambda}\right)^{\ell}=exp\left(\frac{h\lambda}{1-\frac{1}{2}h\lambda}\right).$$

Therefore,

$$\|e_n\| \leq \frac{c}{\lambda}h^2\left[\left(\frac{1+\frac{1}{2}h\lambda}{1-\frac{1}{2}h\lambda}\right)^n-1\right]\leq\frac{c}{\lambda}h^2\left(\frac{1+\frac{1}{2}h\lambda}{1-\frac{1}{2}h\lambda}\right)^n\leq\frac{c}{\lambda}h^2exp\left(\frac{nh\lambda}{1-\frac{1}{2}h\lambda}\right). \tag{169}$$

This bound of true for every negative integer n such that $nh<T$. Therefore,

$$\|e_n\| \leq \frac{c}{\lambda}h^2exp\left(\frac{nh\lambda}{1-\frac{1}{2}h\lambda}\right)\leq\frac{c}{\lambda}h^2exp\left(\frac{T\lambda}{1-\frac{1}{2}h\lambda}\right). \tag{170}$$

So, from the claim (170), we get $\|e_n\|\to 0$, when $h\to 0$. Therefore Trapezoidal Method is convergent . □

Definition 6.8. *(tableaux) The tableaux of Trapezoidal method*

0	0	0
1	$\frac{1}{2}$	$\frac{1}{2}$
	$\frac{1}{2}$	$\frac{1}{2}$.

6.3 Theta Method

Definition 6.9. *(Theta Method[a])*

$$y_{n+1}=y_n+h[\theta f(t_n,y_n)+(1-\theta)f(t_{n+1},y_{n+1})], \quad n=0,1,2,\cdots. \tag{171}$$

[a] Theta Method is a general form of Forward Euler Method ($\theta=1$), Backward Euler Method ($\theta=0$) and Trapezoidal Method ($\theta=\frac{1}{2}$).

Definition 6.10. *(tableaux) The tableaux of θ-method*

0	0	0
1	θ	1-θ
	θ	1-θ.

Solution. Since, the θ-Method's scheme is as follows,

$$y_{n+1} = y_n + h[\theta f(t_n, y_n) + (1-\theta) f(t_{n+1}, y_{n+1})], \quad n = 0, 1, 2, \cdots.$$

. Then, this scheme can be rewritten as RK-scheme, i.e.

$$\begin{aligned} \xi_1 &= y_n \\ \xi_2 &= y_n + h[\theta f(t_n + 0h, \xi_1) + (1-\theta)(t_n + 1h, \xi_2)] \\ y_{n+1} &= y_n + h[\theta f(t_n + 0h, \xi_1) + (1-\theta) f(t_n + h, \xi_2)] \end{aligned}$$

So, the tableaux of θ-method is

0	0	0
1	θ	1-θ
	θ	1-θ.

◀

6.4 Midpoint Rule Method

Definition 6.11. *(Midpoint Rule Method)*

$$y_{n+1} = y_n + hf\left(t_n + \frac{1}{2}h, \frac{1}{2}(y_n + y_{n+1})\right). \tag{172}$$

Theorem 6.7. *(Midpoint Rule Method is of order 2) Midpoint Rule Method*

$$y(t_{n+1}) = y(t_n) + hf\left(t_n + \frac{1}{2}h, \frac{1}{2}(y(t_n) + y(t_{n+1}))\right). \tag{173}$$

is of order 2 .

Proof. By the Taylor expansion,

$$y(t_{n+1}) = y(t_n) + hy'(t_n) + \frac{1}{2!}h^2 y''(t_n) + \mathcal{O}(h^3) \tag{174}$$

$$f(x_0 + \Delta x, y_0 + \Delta y) = f(x_0, y_0) + \left(\Delta x \frac{\partial}{\partial x} + \Delta y \frac{\partial}{\partial y}\right) f(x_0, y_0) + \mathcal{O}(h^2). \tag{175}$$

And chain rule

$$y'' = f'(t, \mathbf{y}) = \frac{\partial f(t, \mathbf{y})}{\partial t} + \frac{\partial f(t, \mathbf{y})}{\partial \mathbf{y}} f(t, \mathbf{y}). \tag{176}$$

So,

$$\begin{aligned} & y(t_{n+1}) - y(t_n) + hf\left(t_n + \frac{1}{2}h, \frac{1}{2}(y(t_n) + y(t_{n+1}))\right) \\ = \; & y(t_n) + hy'(t_n) + \frac{1}{2!}h^2 y''(t_n) + \mathcal{O}(h^3) - y(t_n) \\ - \; & h\left(f(t_n, y_n) + (t_n + \frac{1}{2}h - t_n)\frac{\partial f(t_n, y_n)}{\partial t} + (\frac{1}{2}(y(t_n) + y(t_{n+1})) - y_n)\frac{\partial f(t_n, y_n)}{\partial y} + \mathcal{O}(h^2)\right) \end{aligned}$$

$$
\begin{aligned}
&= hy'(t_n) + \frac{1}{2!}h^2 y''(t_n) + \mathcal{O}(h^3) \\
&- \left(hf(t_n, y_n) + \frac{1}{2}h^2 \frac{\partial f(t_n, y_n)}{\partial t} + \frac{1}{2}h^2 \frac{\partial f(t_n, y_n)}{\partial y} + \mathcal{O}(h^3) \right) \\
&= hy'(t_n) + \frac{1}{2!}h^2 \left(\frac{\partial f(t_n, y_n)}{\partial t} + \frac{\partial f(t_n, y_n)}{\partial y} y'(t_n) \right) \\
&- \left(hy'(t_n) + \frac{1}{2}h^2 \frac{\partial f(t_n, y_n)}{\partial t} + \frac{1}{2}h^2 \frac{\partial f(t_n, y_n)}{\partial y} + \mathcal{O}(h^3) \right) \\
&= \mathcal{O}(h^3).
\end{aligned}
$$

Therefore, Midpoint Rule Method (6.7) is order of 2 . □

Theorem 6.8. *(The convergence of Midpoint Rule Method) Midpoint Rule Method*

$$y(t_{n+1}) = y(t_n) + hf\left(t_n + \frac{1}{2}h, \frac{1}{2}(y(t_n) + y(t_{n+1}))\right), \tag{177}$$

is convergent.

Proof. From (177), we get

$$y(t_{n+1}) = y(t_n) + hf\left(t_n + \frac{1}{2}h, \frac{1}{2}(y(t_n) + y(t_{n+1}))\right) + \mathcal{O}(h^3), \tag{178}$$

Subtracting (178) from (172), we get

$$e_{n+1} = e_n + h\left[f\left(t_n + \frac{1}{2}h, \frac{1}{2}(y(t_n) + y(t_{n+1}))\right) - f\left(t_n + \frac{1}{2}h, \frac{1}{2}(y(t_n) + y(t_{n+1}))\right)\right] + ch^3. \tag{179}$$

Since f is lipschitz continuous w.r.t. the second variable, then

$$
\begin{aligned}
&\left| f\left(t_n + \frac{1}{2}h, \frac{1}{2}(y(t_n) + y(t_{n+1}))\right) - f\left(t_n + \frac{1}{2}h, \frac{1}{2}(y(t_n) + y(t_{n+1}))\right)\right| \\
\leq\ & \frac{1}{2}\lambda |y_n - y(t_n) + y_{n+1} - y(t_{n+1})|, \quad \lambda > 0.
\end{aligned} \tag{180}
$$

Therefore,

$$\left\|e_{n+1}\right\| \leq \|e_n\| + \frac{1}{2}h\lambda(\|e_n\| + \left\|e_{n+1}\right\|) + ch^3. \tag{181}$$

So,

$$(1 - \frac{1}{2}h\lambda)\left\|e_{n+1}\right\| \leq (1 + \frac{1}{2}h\lambda)\|e_n\| + ch^3. \tag{182}$$

Claim:[2]

$$\|e_n\| \leq \frac{c}{\lambda}h^2\left[\left(\frac{1 + \frac{1}{2}h\lambda}{1 - \frac{1}{2}h\lambda}\right)^n - 1\right], n = 0, 1, \cdots \tag{183}$$

Proof for Claim (183): The proof is by induction on n.

Then, we can make h small enough to such that $0 < h\lambda < 2$, then

$$\frac{1+\frac{1}{2}h\lambda}{1-\frac{1}{2}h\lambda} = 1 + \frac{1}{1-\frac{1}{2}h\lambda} \leq \sum_{\ell=0}^{\infty} \frac{1}{\ell!}\left(\frac{h\lambda}{1-\frac{1}{2}h\lambda}\right)^{\ell} = exp\left(\frac{h\lambda}{1-\frac{1}{2}h\lambda}\right).$$

Therefore,

$$\|e_n\| \leq \frac{c}{\lambda}h^2\left[\left(\frac{1+\frac{1}{2}h\lambda}{1-\frac{1}{2}h\lambda}\right)^n - 1\right] \leq \frac{c}{\lambda}h^2\left(\frac{1+\frac{1}{2}h\lambda}{1-\frac{1}{2}h\lambda}\right)^n \leq \frac{c}{\lambda}h^2 exp\left(\frac{nh\lambda}{1-\frac{1}{2}h\lambda}\right). \tag{184}$$

This bound of true for every negative integer n such that $nh < T$. Therefore,

$$\|e_n\| \leq \frac{c}{\lambda}h^2 exp\left(\frac{nh\lambda}{1-\frac{1}{2}h\lambda}\right) \leq \frac{c}{\lambda}h^2 exp\left(\frac{T\lambda}{1-\frac{1}{2}h\lambda}\right). \tag{185}$$

So, from the claim (185), we get $\|e_n\| \to 0$, when $h \to 0$. Therefore Midpoint Rule Method is convergent . □

6.5 Problems

Problem 6.1. *(Prelim Aug. 2013#1)*

Solution. ◄

7 Multistep Methond

7.1 The Adams Method

Definition 7.1. *(s-step Adams-bashforth)*

$$y_{n+s} = y_{n+s-1} + h\sum_{m=0}^{s-1} b_m f(t_{n+m}, y_{n+m}), \tag{186}$$

where

$$\begin{aligned} b_m &= h^{-1}\int_{t_{n+s-1}}^{t_{n+s}} p_m(\tau)d\tau = h^{-1}\int_0^h p_m(t_{n+s-1}+\tau)d\tau \quad n = 0,1,2,\cdots. \\ p_m(t) &= \Pi_{l=0,l\neq m}^{s-1}\frac{t-t_{n+l}}{t_{n+m}-t_{n+l}}, \quad \textit{Lagrange interpolation polynomials}. \end{aligned}$$

(1-step Adams-bashforth)

$$y_{n+1} = y_n + hf(t_n, y_n),$$

(2-step Adams-bashforth)

$$y_{n+2} = y_{n+1} + h\left[\frac{3}{2}f(t_{n+1}, y_{n+1}) - \frac{1}{2}f(t_n, y_n)\right],$$

(3-step Adams-bashforth)

$$y_{n+3} = y_{n+2} + h\left[\frac{23}{12}f(t_{n+2}, y_{n+2}) - \frac{4}{3}f(t_{n+1}, y_{n+1}) + \frac{5}{12}f(t_n, y_n)\right].$$

7.2 The Order and Convergence of Multistep Methods

Definition 7.2. *(General s-step Method) The general s-step Method* [a] *can be written as*

$$\sum_{m=0}^{s} a_m \boldsymbol{y}_{n+m} = h\sum_{m=0}^{s} b_m \boldsymbol{f}(t_{n+m}, \boldsymbol{y}_{n+m}). \tag{187}$$

Where $a_m, b_m, m = 0,\cdots,s$, are given constants, independent of h, n and original equation.

[a] if $b_s = 0$ the method is explicit; otherwise it is implicit.

Theorem 7.1. *(s-step method convergent order) The multistep method (187) is of order $p \geq 1$ if and only if there exists $c \neq 0$ s.t.*

$$\rho(w) - \sigma(w)\ln w^{a} = c(w-1)^{p+1} + \mathcal{O}(|w-1|^{p+2}), \quad w \to 1. \tag{188}$$

Where,

$$\rho(w) := \sum_{m=0}^{s} a_m w^m \quad \text{and} \quad \sigma(w) := \sum_{m=0}^{s} b_m w^m. \tag{189}$$

[a] Let $w = \xi + 1$, then $\ln(1+\xi) = \sum_{n=0}^{\infty}(-1)^n \frac{\xi^{n+1}}{n+1} = \xi - \frac{\xi^2}{2} + \frac{\xi^3}{3} - \frac{\xi^4}{4} + \cdots + (-1)^n \frac{\xi^{n+1}}{n+1} + \cdots, \xi \in (-1,1)$.

Theorem 7.2. *(s-step method convergent order) The multistep method (187) is of order $p \geq 1$ if and only if*

1. $\sum_{m=0}^{s} a_m = 0$, $(i.e. \rho(1) = 0)$,
2. $\sum_{m=0}^{s} m^k a_m = k \sum_{m=0}^{s} m^{k-1} b_m, k = 1, 2, \cdots, p$,
3. $\sum_{m=0}^{s} m^{p+1} a_m \neq (p+1) \sum_{m=0}^{s} m^p b_m$.

Where,

$$\rho(w) := \sum_{m=0}^{s} a_m w^m \quad \text{and} \quad \sigma(w) := \sum_{m=0}^{s} b_m w^m. \tag{190}$$

Lemma 7.1. *(Root Condition) If the roots $|\lambda_i| \leq 1$ for each $i = 1, \cdots, m$ and all roots with value 1 are simple root then the difference method is said to satisfy the root condition.*

Theorem 7.3. *(The Dahlquist equivalence theorem) The multistep method (187) is convergent if and only if*

1. *consistency: multistep method (187)is order of $p \geq 1$,*
2. *stability: the polynomial $\rho(w)$ satisfies the root condition .*

7.3 Method of A-stable verification for Multistep Methods

Theorem 7.4. *Explicit Multistep Methods can not be A-stable.*

Theorem 7.5. *(Dahlquist second barrier) The highest oder of an A-stable multistep method is 2 .*

7.4 Problems

Problem 7.1. *Find the order of the following quadrature formula.*

$$\int_0^1 f(\tau)d\tau = \frac{1}{6}f(0) + \frac{2}{3}f(\frac{1}{2}) + \frac{1}{6}f(1), \quad \textit{Simpson Rule.}$$

Solution. Since the quadrature formula (209) is order of p if it is exact for every $f \in \mathbb{P}_{p-1}$. we can chose

the simplest basis $(1,\tau,\tau^2,\tau^3,\cdots,\tau^{p-1})$, and the order conditions read that

$$\sum_{j=1}^{p} b_j c_j^m = \int_a^b \tau^m w(\tau)d\tau, \quad m=0,1,\cdots,p-1. \tag{191}$$

Checking the order condition by the following procedure,

$$\begin{aligned}
1=\int_0^1 1d\tau &= \frac{1}{6}1+\frac{2}{3}1+\frac{1}{6}1=1.\\
\frac{1}{2}=\int_0^1 \tau d\tau &= \frac{1}{6}0+\frac{2}{3}\left(\frac{1}{2}\right)+\frac{1}{6}1=\frac{1}{2}.\\
\frac{1}{3}=\int_0^1 \tau^2 d\tau &= \frac{1}{6}0^2+\frac{2}{3}\left(\frac{1}{2}\right)^2+\frac{1}{6}1^2=\frac{1}{3}.\\
\frac{1}{4}=\int_0^1 \tau^3 d\tau &= \frac{1}{6}0^3+\frac{2}{3}\left(\frac{1}{2}\right)^3+\frac{1}{6}1^3=\frac{1}{4}.\\
\frac{1}{5}=\int_0^1 \tau^4 d\tau &\neq \frac{1}{6}0^4+\frac{2}{3}\left(\frac{1}{2}\right)^4+\frac{1}{6}1^4=\frac{5}{24}.
\end{aligned}$$

we can get the order of the Simpson rule quadrature formula is 4. ◀

Problem 7.2. *Recall Simpson's quadrature rule:*

$$\int_a^b f(\tau)d\tau=\frac{b-a}{6}\left[f(a)+4f(\frac{a+b}{2})+f(b)\right]+\mathcal{O}(|b-a|^4), \quad \textit{Simpson Rule.}$$

Starting from the identity

$$y(t_{n+1})-y(t_{n-1})=\int_{t_{n-1}}^{t_{n+1}} f(s;y(s))ds. \tag{192}$$

use Simpson's rule to derive a 3-step method. Determine its order and whether it is convergent.

Solution. 1. The derivation of the a 3-step method

since,

$$y(t_{n+1})-y(t_{n-1})=\int_{t_{n-1}}^{t_{n+1}} f(s;y(s))ds. \tag{193}$$

Then,by Simpson's quadrature rule, we have

$$\begin{aligned}
& y(t_{n+1})-y(t_{n-1}) && (194)\\
=\ & \int_{t_{n-1}}^{t_{n+1}} f(s;y(s))ds. && (195)\\
=\ & \frac{t_{n+1}-t_{n-1}}{6}\left[f(t_{n-1};y(t_{n-1}))+4f\left(\frac{t_{n-1}+t_{n+1}}{2};y\left(\frac{t_{n-1}+t_{n+1}}{2}\right)\right)+f(t_{n+1};y(t_{n+1}))\right] && (196)\\
=\ & \frac{h}{3}\left[f(t_{n-1};y(t_{n-1}))+4f\left(t_n;y\left(t_n\right)\right)+f(t_{n+1};y(t_{n+1}))\right]. && (197)
\end{aligned}$$

Therefore, the 3-step method deriving from Simpson's rule is

$$y(t_{n+1}) = y(t_{n-1}) + \frac{h}{3}\left[f(t_{n-1}; y(t_{n-1})) + 4f(t_n; y(t_n)) + f(t_{n+1}; y(t_{n+1}))\right]. \tag{198}$$

Or

$$y(t_{n+2}) - y(t_n) = \frac{h}{3}\left[f(t_n; y(t_n)) + 4f(t_{n+1}; y(t_{n+1})) + f(t_{n+2}; y(t_{n+2}))\right]. \tag{199}$$

2. The order For our this problem

$$\rho(w) := \sum_{m=0}^{s} a_m w^m = -1 + w^2 \quad \text{and} \quad \sigma(w) := \sum_{m=0}^{s} b_m w^m = \frac{1}{3} + \frac{4}{3}w + \frac{1}{3}w^2. \tag{200}$$

By making the substitution with $\xi = w - 1$ i.e. $w = \xi + 1$, then

$$\rho(w) := \sum_{m=0}^{s} a_m w^m = \xi^2 + 2\xi \quad \text{and} \quad \sigma(w) := \sum_{m=0}^{s} b_m w^m = \frac{1}{3}\xi^2 + 2\xi + 2. \tag{201}$$

So,

$$\begin{aligned}
\rho(w) - \sigma(w) ln(w) &= \xi^2 + 2\xi - (2 + 2\xi + \frac{1}{3}\xi^2)(\xi - \frac{\xi^2}{2} + \frac{\xi^3}{3} \cdots) \\
&= \begin{array}{ccccc} +2\xi & +\xi^2 & & & \\ -2\xi & +\xi^2 & -\frac{2}{3}\xi^3 & & \\ & -2\xi^2 & +\xi^3 & -\frac{2}{3}\xi^4 & \\ & & -\frac{1}{3}\xi^3 & +\frac{1}{6}\xi^4 & -\frac{1}{9}\xi^5 \end{array} \\
&= -\frac{1}{2}\xi^4 + \mathcal{O}(\xi^5).
\end{aligned}$$

Therefore, by the theorem

$$\rho(w) - \sigma(w) ln(w) = -\frac{1}{2}\xi^4 + \mathcal{O}(\xi^5).$$

Hence, this scheme is order of 3.

3. The stability Since,

$$\rho(w) := \sum_{m=0}^{s} a_m w^m = -1 + w^2 = (w-1)(w+1). \tag{202}$$

And $w = \pm 1$ are simple root which satisfy the root condition. Therefore, this scheme is stable. Hence, it is of order 3 and convergent. convergent ◀

Problem 7.3. *Restricting your attention to scalar autonomous $y' = f(y)$, prove that the ERK method with tableau*

$$\begin{array}{c|cccc} 0 & & & & \\ \frac{1}{2} & \frac{1}{2} & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & & \\ 1 & 0 & 0 & 1 & \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array}$$

is of order 4.

Solution. ◀

Problem 7.4. *(Prelim Jan. 2011#5) Consider*

$$y'(t) = f(t, y(t)), \quad t \geq t_0, y(t_0) = y_0,$$

where $f : [t_0, t^*] \times \mathbb{R} \to \mathbb{R}$ is continuous in its first variable and Lipschitz continuous in its second variable. Prove that Euler's method converges.

Solution. The Euler's scheme is as follows:

$$y_{n+1} = y_n + hf(t_n, y_n), \quad n = 0, 1, 2, \cdots. \tag{203}$$

By the Taylor expansion,

$$y(t_{n+1}) = y(t_n) + hy'(t_n) + \mathcal{O}(h^2).$$

So,

$$\begin{aligned} y(t_{n+1}) - y(t_n) - hf(t_n, y(t_n)) &= y(t_n) + hy'(t_n) + \mathcal{O}(h^2) - y(t_n) - hf(t_n, y(t_n)) \\ &= y(t_n) + hy'(t_n) + \mathcal{O}(h^2) - y(t_n) - hy'(t_n) \\ &= \mathcal{O}(h^2). \end{aligned} \tag{204}$$

Therefore, Forward Euler Method is order of 1 .

From (219), we get

$$y(t_{n+1}) = y(t_n) + hf(t_n, y(t_n)) + \mathcal{O}(h^2), \tag{205}$$

Subtracting (220) from (218), we get

$$e_{n+1} = e_n + h[f(t_n, y_n) - f(t_n, y(t_n))] + ch^2.$$

Since f is lipschitz continuous w.r.t. the second variable, then

$$|f(t_n, y_n) - f(t_n, y(t_n))| \leq \lambda |y_n - y(t_n)|, \quad \lambda > 0.$$

Therefore,

$$\begin{aligned} \|e_{n+1}\| &\leq \|e_n\| + h\lambda \|e_n\| + ch^2 \\ &= (1 + h\lambda)\|e_n\| + ch^2. \end{aligned}$$

Claim:[2]

$$\|e_n\| \leq \frac{c}{\lambda} h[(1 + h\lambda)^n - 1], n = 0, 1, \cdots$$

Proof for Claim (221): The proof is by induction on n.

1. when $n = 0$, $e_n = 0$, hence $\|e_n\| \leq \frac{c}{\lambda} h[(1 + h\lambda)^n - 1]$,
2. Induction assumption:

$$\|e_n\| \leq \frac{c}{\lambda} h[(1 + h\lambda)^n - 1]$$

3. Induction steps:

$$\begin{aligned}\|e_{n+1}\| &\leq (1+h\lambda)\|e_n\| + ch^2 \\ &\leq (1+h\lambda)\frac{c}{\lambda}h[(1+h\lambda)^n - 1] + ch^2 \\ &= \frac{c}{\lambda}h[(1+h\lambda)^{n+1} - 1].\end{aligned}$$

So, from the claim (221), we get $\|e_n\| \to 0$, when $h \to 0$. Therefore Forward Euler Method is convergent .

◀

Problem 7.5. *(Prelim Jan. 2011#6) Consider the scheme*

$$y_{n+2} + y_{n+1} - 2y_n = h(f(t_{n+2}, y_{n+2}) + f(t_{n+1}, y_{n+1}) + f(t_n, y_n))$$

for approximating the solution to

$$y'(t) = f(t, y(t)), \quad t \geq t_0, y(t_0) = y_0,$$

what's the order of the scheme? Is it a convergent scheme? Is it A-stable? Justify your answers.

Solution. For our this problem

$$\rho(w) := \sum_{m=0}^{s} a_m w^m = -2 + w + w^2 \quad \text{and} \quad \sigma(w) := \sum_{m=0}^{s} b_m w^m = 1 + w + w^2. \tag{206}$$

By making the substitution with $\xi = w - 1$ i.e. $w = \xi + 1$, then

$$\rho(w) := \sum_{m=0}^{s} a_m w^m = \xi^2 + 3\xi \quad \text{and} \quad \sigma(w) := \sum_{m=0}^{s} b_m w^m = \xi^2 + 3\xi + 3. \tag{207}$$

So,

$$\begin{aligned}\rho(w) - \sigma(w)ln(w) &= \xi^2 + 3\xi - (3 + 3\xi + \xi^2)(\xi - \frac{\xi^2}{2} + \frac{\xi^3}{3}\cdots) \\ &= \begin{array}{cccc} +3\xi & +\xi^2 & & \\ -3\xi & -3\xi^2 & -\xi^3 & \\ & +\frac{3}{2}\xi^2 & +\frac{3}{2}\xi^3 & +\frac{1}{2}\xi^4 \\ & & -\xi^3 & -\xi^4 \quad -\frac{1}{3}\xi^5 \end{array} \\ &= -\frac{1}{2}\xi^2 + \mathcal{O}(\xi^3).\end{aligned}$$

Therefore, by the theorem

$$\rho(w) - \sigma(w)ln(w) = -\frac{1}{2}\xi^2 + \mathcal{O}(\xi^3).$$

Hence, this scheme is order of 1. The stability Since,

$$\rho(w) := \sum_{m=0}^{s} a_m w^m = -2 + w + w^2 = (w+2)(w-1). \tag{208}$$

And $w = -1$ or $w = -2$ which does not satisfy the root condition. Therefore, this scheme is not stable. Hence, it is also not A-stable.

◀

Problem 7.6. *(Prelim Jan. 2011#4)*

Solution. ◀

8 Runge-Kutta Methods

8.1 Quadrature Formulas

Definition 8.1. *(The Quadrature) The Quadrature is the procedure of replacing an integral with a finite sum.*

Definition 8.2. *(The Quadrature Formula) Let w be a nonnegative function in (a,b) s.t.*

$$0 < \int_a^b w(\tau)d\tau < \infty, \quad \left|\int_a^b \tau^j w(\tau)d\tau\right| < \infty, j = 1,2,\cdots.$$

Then, the quadrature formula is as following

$$\int_a^b f(\tau)w(\tau)d\tau \approx \sum_j^n b_j f(c_j). \tag{209}$$

Remark 8.1. *The quadrature formula (209) is order of p if it is exact for every $f \in \mathbb{P}_{p-1}$.*

8.2 Explicit Runge-Kutta Formulas

Definition 8.3. *(Explicit Runge-Kutta Formulas) Explicit Runge-Kutta is to integral from t_n to t_{n+1} as follows*

$$\begin{aligned} y(t_{n+1}) &= y(t_n) + \int_{t_n}^{t_{n+1}} f(\tau, y(\tau))d\tau \\ &= y(t_n) + h\int_0^1 f(t_n + h\tau, y(t_n + h\tau))d\tau \end{aligned}$$

and to replace the second integral by a quadrature, i.e.

$$y_{n+1} = y_n + h\sum_{j=1}^{\nu} b_j f(t_n + c_j h, y(t_n + c_j h))$$

Specifically, we have

$$\begin{aligned} \xi_1 &= y_n, \\ \xi_2 &= y_n + ha_{21}f(t_n, \xi_1) \\ \xi_3 &= y_n + ha_{31}f(t_n + c_1 h, \xi_1) + ha_{32}f(t_n + c_2 h, \xi_2) \\ &\vdots \\ \xi_\nu &= y_n + h\sum_{i=1}^{\nu-1} a_{\nu i} f(t_n + c_i h, \xi_i)) \\ y_{n+1} &= y_n + h\sum_{j=1}^{\nu} b_j f(t_n + c_j h, \xi_j)). \end{aligned}$$

Definition 8.4. *(tableaux) The tableaux of REK*

$$\begin{array}{c|c} c & A \\ \hline & b^T \end{array}$$

where A is low triangular matrix.

Remark 8.2. *by observing that the condition*

$$\sum_{i=1}^{j-1} a_{j,i} = c_j, \quad j = 2,3,\cdots,\nu,$$

is necessary for order 1.

8.3 Implicit Runge-Kutta Formulas

Definition 8.5. *(Implicit Runge-Kutta Formulas) Implicit Runge-Kutta use the following scheme*

$$\begin{aligned} \xi_j &= y_n + h\sum_{i=1}^{\nu} a_{j,i} f(t_n + c_i h, \xi_i), j = 1,2,\cdots,\nu \\ y_{n+1} &= y_n + h\sum_{j=1}^{\nu} b_j f(t_n + c_j h, \xi_j). \end{aligned}$$

8.4 Method of A-stable verification for Runge-Kutta Method

Theorem 8.1. *Explicit Runge-Kutta Methods can not be A-stable.*

Theorem 8.2. *necessary & sufficient A necessary & sufficient condition for A-stable Runge-Kutta method is*

$$\left|r(z)\right| < 1, \ z \in \mathbb{C}^-,$$

where

$$r(z) = 1 + zb^T(I - zA)^{-1}\mathbb{1}.$$

8.5 Problems

9 Finite Difference Method

Definition 9.1. *(Discrete 2-norm) The discrete 2-norm is defined as follows*

$$\|u\|_{2,h}^2 = h^d \sum_{i=1}^{N} |u_i|^2,$$

where d is dimension.

Theorem 9.1. *(Discrete maximum principle) Let $A = tridiag\{a_i, b_i, c_i\}_{i=1}^n \in \mathbb{R}^{n\times n}$ be a tridiagional matrix with the properties that*

$$b_i > 0, \;\; a_i, c_i \le 0, \;\; a_i + b_i + c_i = 0.$$

Prove the following maximum principle: If $u \in \mathbb{R}^n$ is such that $(Au)_{i=2,\cdots,n-1} \le 0$, then $u_i \le \max\{u_1, u_n\}$.

Proof. Without loss generality, we assume $u_k, k = 2, \cdots, n-1$ is the maximum value.

1. For $(Au)_{i=2,\cdots,n-1} < 0$:
   I will use the method of contradiction to prove this case. Since $(Au)_{i=2,\cdots,n-1} < 0$, so

   $$a_k u_{k-1} + b_k u_k + c_k u_{k+1} < 0.$$

   Since $a_k + c_k = -b_k$ and $a_k < 0, c_k < 0$, so

   $$a_k u_{k-1} - (a_k + c_k)u_k + c_k u_{k+1} = a_k(u_{k-1} - u_k) + c_k(u_{k+1} - u_k) \ge 0.$$

   This is contradiction to $(Au)_{i=2,\cdots,n-1} < 0$. Therefore, If $u \in \mathbb{R}^n$ is such that $(Au)_{i=2,\cdots,n-1} < 0$, then $u_i < \max\{u_1, u_n\}$.
2. For $(Au)_{i=2,\cdots,n-1} = 0$:
   Since $(Au)_{i=2,\cdots,n-1} = 0$, so

   $$a_k u_{k-1} + b_k u_k + c_k u_{k+1} = 0.$$

   Since $a_k + c_k = -b_k$, so

   $$a_k u_{k-1} - (a_k + c_k)u_k + c_k u_{k+1} = a_k(u_{k-1} - u_k) + c_k(u_{k+1} - u_k) = 0.$$

   And $a_k < 0, c_k < 0, u_{k-1} - u_k \le 0, u_{k+1} - u_k \le 0$, so $u_{k-1} = u_k = u_{k+1}$, that is to say, u_{k-1} and u_{k+1} is also the maximum points. Bu using the same argument again, we get $u_{k-2} = u_{k-1} = u_k = u_{k+1} = u_{k+2}$. Repeating the process, we get

   $$u_1 = u_2 = \cdots = u_{n-1} = u_n.$$

   Therefore, If $u \in \mathbb{R}^n$ is such that $(Au)_{i=2,\cdots,n-1} = 0$, then $u_i \le \max\{u_1, u_n\}$

□

Theorem 9.2. *(Discrete Poincaré inequality) Let $\Omega = (0,1)$ and Ω_h be a uniform grid of size h. If $Y \in \mathcal{U}_h$ is a mesh function on Ω_h such that $Y(0) = 0$, then there is a constant C, independent of Y and h, for which*

$$\|Y\|_{2,h} \le C \left\|\delta Y\right\|_{2,h}.$$

Proof. I consider the following uniform partition (Figure. A1) of the interval $(0,1)$ with N points.

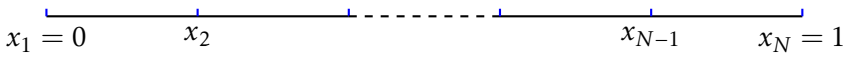

Figure 3: One dimension's uniform partition

Since the discrete 2-norm is defined as follows

$$\|v\|_{2,h}^2 = h^d \sum_{i=1}^{N} |v_i|^2,$$

where d is dimension. So, we have

$$\|v\|_{2,h}^2 = h\sum_{i=1}^{N} |v_i|^2, \quad \left\|\bar{\delta}v\right\|_{2,h}^2 = h\sum_{i=2}^{N} \left|\frac{v_{i-1}-v_i}{h}\right|^2.$$

Since $Y(0) = 0$, i.e. $Y_1 = 0$,

$$\sum_{i=2}^{N} (Y_{i-1} - Y_i) = Y_1 - Y_N = -Y_N.$$

Then,

$$\left|\sum_{i=2}^{N} (Y_{i-1} - Y_i)\right| = |Y_N|.$$

and

$$|Y_N| \le \sum_{i=2}^{N} |Y_{i-1} - Y_i| = \sum_{i=2}^{N} h\left|\frac{Y_{i-1}-Y_i}{h}\right| \le \left(\sum_{i=2}^{N} h^2\right)^{1/2} \left(\sum_{i=2}^{N} \left|\frac{Y_{i-1}-Y_i}{h}\right|^2\right)^{1/2}.$$

Therefore

$$\begin{aligned} |Y_K|^2 &\le \left(\sum_{i=2}^{K} h^2\right)\left(\sum_{i=2}^{K} \left|\frac{Y_{i-1}-Y_i}{h}\right|^2\right) \\ &= h^2(K-1)\sum_{i=2}^{K} \left|\frac{Y_{i-1}-Y_i}{h}\right|^2. \end{aligned}$$

1. When $K = 2$,

$$|Y_2|^2 \le h^2 \left|\frac{Y_1 - Y_2}{h}\right|^2.$$

2. When $K = 3$,

$$|Y_3|^2 \le 2h^2 \left(\left|\frac{Y_1 - Y_2}{h}\right|^2 + \left|\frac{Y_2 - Y_3}{h}\right|^2\right).$$

3. When $K = N$,

$$|Y_N|^2 \leq (N-1)h^2\left(\left|\frac{Y_1 - Y_2}{h}\right|^2 + \left|\frac{Y_2 - Y_3}{h}\right|^2 + \cdots + \left|\frac{Y_{N-1} - Y_N}{h}\right|^2\right).$$

Sum over $|Y_i|^2$ from 2 to N, we get

$$\sum_{i=2}^{N} |Y_i|^2 \leq \frac{N(N-1)}{2} h^2 \sum_{i=2}^{N} \left|\frac{Y_{i-1} - Y_i}{h}\right|^2.$$

Since $Y_1 = 0$, so

$$\sum_{i=1}^{N} |Y_i|^2 \leq \frac{N(N-1)}{2} h^2 \sum_{i=2}^{N} \left|\frac{Y_{i-1} - Y_i}{h}\right|^2.$$

And then

$$\frac{1}{(N-1)^2} \sum_{i=1}^{N} |Y_i|^2 \leq \frac{N}{2(N-1)} h^2 \sum_{i=2}^{N} \left|\frac{Y_{i-1} - Y_i}{h}\right|^2 = \left(\frac{1}{2} + \frac{1}{2(N-1)}\right) h^2 \sum_{i=2}^{N} \left|\frac{Y_{i-1} - Y_i}{h}\right|^2.$$

Since $h = \frac{1}{N-1}$, so

$$h^2 \sum_{i=1}^{N} |Y_i|^2 \leq \left(\frac{1}{2} + \frac{1}{2(N-1)}\right) h^2 \sum_{i=2}^{N} \left|\frac{Y_{i-1} - Y_i}{h}\right|^2.$$

then

$$h \sum_{i=1}^{N} |Y_i|^2 \leq \left(\frac{1}{2} + \frac{1}{2(N-1)}\right) h \sum_{i=2}^{N} \left|\frac{Y_{i-1} - Y_i}{h}\right|^2.$$

i.e,

$$\|Y\|_{2,h}^2 \leq \left(\frac{1}{2} + \frac{1}{2(N-1)}\right) \left\|\bar{\delta} Y\right\|_{2,h}^2.$$

since $N \geq 2$, so

$$\|Y\|_{2,h}^2 \leq \left\|\bar{\delta} Y\right\|_{2,h}^2.$$

Hence,

$$\|Y\|_{2,h} \leq C \left\|\bar{\delta} Y\right\|_{2,h}.$$

□

Theorem 9.3. *(Von Neumann stability analysis method) For the difference scheme*

$$U_j^{n+1} = \sum_{p \in \mathbb{N}} \alpha_p U_{j-p}^n,$$

we have the corresponding Fourier transform is as follows

$$\hat{U}^{n+1}(\xi) = \sum_{p \in \mathbb{N}} \alpha_p e^{-ip\xi} \hat{U}^n(\xi) := G(\lambda, \xi) \hat{U}^n(\xi).$$

Where $\lambda = \frac{\tau}{h^2}$ is CFL number and $G(\lambda, \xi)$ is called Growth factor . If $\left|G(\lambda, \xi)\right| \leq 1$, then the difference scheme is stable.

9.1 Problems

Problem 9.1. *(Prelim Jan. 2011#7) Consider the Crank-Nicholson scheme applied to the diffusion equation*

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$$

where $t > 0, -\infty < x < \infty$.

1. *Show that the amplification factor in the Von Neumann analysis of the scheme us*

$$g(\xi) = \frac{1+\frac{1}{2}z}{1-\frac{1}{2}z}, z = 2\frac{\Delta t}{\Delta x^2}(cos(\Delta x\xi) - 1).$$

2. *Use the results of part 1 to show that the scheme is stable.*

Solution. 1. The Crank-Nicholson scheme for the diffusion equation is

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} = \frac{1}{2}\left(\frac{u_{j-1}^{n+1} - 2u_j^{n+1} + u_{j+1}^{n+1}}{\Delta x^2} + \frac{u_{j-1}^n - 2u_j^n + u_{j+1}^n}{\Delta x^2}\right)$$

Let $\mu = \frac{\Delta t}{\Delta x^2}$, then the scheme can be rewrote as

$$u_j^{n+1} = u_j^n + \frac{\mu}{2}\left(u_{j-1}^{n+1} - 2u_j^{n+1} + u_{j+1}^{n+1} + u_{j-1}^n - 2u_j^n + u_{j+1}^n\right),$$

i.e.

$$-\frac{\mu}{2}u_{j-1}^{n+1} + (1+\mu)u_j^{n+1} - \frac{\mu}{2}u_{j+1}^{n+1} = \frac{\mu}{2}u_{j-1}^n + (1-\mu)u_j^n + \frac{\mu}{2}u_{j+1}^n.$$

By using the Fourier transform, i.e.

$$u_j^{n+1} = g(\xi)u_j^n, \;\; u_j^n = e^{ij\Delta x\xi},$$

then we have

$$-\frac{\mu}{2}g(\xi)u_{j-1}^n + (1+\mu)g(\xi)u_j^n - \frac{\mu}{2}g(\xi)u_{j+1}^n = \frac{\mu}{2}u_{j-1}^n + (1-\mu)u_j^n + \frac{\mu}{2}u_{j+1}^n.$$

And then

$$-\frac{\mu}{2}g(\xi)e^{i(j-1)\Delta x\xi} + (1+\mu)g(\xi)e^{ij\Delta x\xi} - \frac{\mu}{2}g(\xi)e^{i(j+1)\Delta x\xi} = \frac{\mu}{2}e^{i(j-1)\Delta x\xi} + (1-\mu)e^{ij\Delta x\xi} + \frac{\mu}{2}e^{i(j+1)\Delta x\xi},$$

i.e.

$$g(\xi)\left(-\frac{\mu}{2}e^{-i\Delta x\xi} + (1+\mu) - \frac{\mu}{2}e^{i\Delta x\xi}\right)e^{j\Delta x\xi} = \left(\frac{\mu}{2}e^{-i\Delta x\xi} + (1-\mu) + \frac{\mu}{2}e^{i\Delta x\xi}\right)e^{j\Delta x\xi},$$

i.e.

$$g(\xi)\left(1+\mu-\mu\cos(\Delta x\xi)\right) = 1-\mu+\mu\cos(\Delta x\xi).$$

therefore,

$$g(\xi) = \frac{1-\mu+\mu\cos(\Delta x\xi)}{1+\mu-\mu\cos(\Delta x\xi)}.$$

hence

$$g(\xi) = \frac{1+\frac{1}{2}z}{1-\frac{1}{2}z}, z = 2\frac{\Delta t}{\Delta x^2}(cos(\Delta x\xi) - 1).$$

2. since $z = 2\frac{\Delta t}{\Delta x^2}(cos(\Delta x\xi) - 1)$, then $z < 0$, then we have

$$1 + \frac{1}{2}z < 1 - \frac{1}{2}z,$$

therefore $g(\xi) < 1$. Since $-1 < 1$, then

$$\frac{1}{2}z - 1 < \frac{1}{2}z + 1.$$

Therefore,

$$g(\xi) = \frac{1 + \frac{1}{2}z}{1 - \frac{1}{2}z} > -1.$$

hence $\left|g(\xi)\right| < 1$. So, the scheme is stable.

◀

Problem 9.2. *(Prelim Jan. 2011#8) Consider the explicit scheme*

$$u_j^{n+1} = u_j^n + \mu\left(u_{j-1}^n - 2u_j^n + u_{j+1}^n\right) - \frac{b\mu\Delta x}{2}\left(u_{j+1}^n - u_{j-1}^n\right), 0 \le n \le N, 1 \le j \le L.$$

for the convention-diffusion problem

$$\begin{cases} \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} - b\frac{\partial u}{\partial x} & \text{for } 0 \le x \le 1,\ 0 \le t \le t^* \\ u(0,t) = u(1,t) = 0 & \text{for } 0 \le t \le t^* \\ u(x,0) = g(x) & \text{for } 0 \le x \le 1, \end{cases}$$

where $b > 0$, $\mu = \frac{\Delta t}{(\Delta x)^2}, \Delta x = \frac{1}{L+1}$, and $\Delta t = \frac{t^*}{N}$. Prove that, under suitable restrictions on μ and Δx, the error grid function e^n satisfy the estimate

$$\|e^n\|_\infty \le t^* C\left(\Delta t + \Delta x^2\right),$$

for all n such that $n\Delta t \le t^*$, where $C > 0$ is a constant.

Solution. Let $\bar{u}$ be the exact solution and $\bar{u}_j^n = \bar{u}(n\Delta t, j\Delta x)$. Then from Taylor Expansion, we have

$$\begin{aligned} \bar{u}_j^{n+1} &= \bar{u}_j^n + \Delta t\frac{\partial}{\partial t}\bar{u}_j^n + \frac{1}{2}(\Delta t)^2\frac{\partial^2}{\partial t^2}\bar{u}(\xi_1, j\Delta x), \quad t_n \le \xi_1 \le t_{n+1}, \\ \bar{u}_{j-1}^n &= \bar{u}_j^n - \Delta x\frac{\partial}{\partial x}\bar{u}_j^n - \frac{1}{6}(\Delta x)^3\frac{\partial^3}{\partial x^3}\bar{u}_j^n + \frac{1}{24}(\Delta x)^4\frac{\partial^4}{\partial x^4}\bar{u}(n\Delta t, \xi_2), \quad x_{j-1} \le \xi_2 \le x_j, \\ \bar{u}_{j+1}^n &= \bar{u}_j^n + \Delta x\frac{\partial}{\partial x}\bar{u}_j^n + \frac{1}{6}(\Delta x)^3\frac{\partial^3}{\partial x^3}\bar{u}_j^n + \frac{1}{24}(\Delta x)^4\frac{\partial^4}{\partial x^4}\bar{u}(n\Delta t, \xi_3), \quad x_j \le \xi_3 \le x_{j+1}. \end{aligned}$$

Then the truncation error T of this scheme is

$$\begin{aligned} T &= \frac{\bar{u}_j^{n+1} - \bar{u}_j^n}{\Delta t} - \frac{\bar{u}_{j-1}^n - 2\bar{u}_j^n + \bar{u}_{j+1}^n}{\Delta x^2} + b\frac{\bar{u}_{j+1}^n - \bar{u}_{j-1}^n}{\Delta x} \\ &= \mathcal{O}(\Delta t + (\Delta x)^2). \end{aligned}$$

Therefore

$$e_j^{n+1} = e_j^n + \mu\left(e_{j-1}^n - 2e_j^n + e_{j+1}^n\right) - \frac{b\mu\Delta x}{2}\left(e_{j+1}^n - e_{j-1}^n\right) + c\Delta t(\Delta t + (\Delta x)^2),$$

i.e.

$$e_j^{n+1} = \left(\mu + \frac{b\mu\Delta x}{2}\right)e_{j-1}^n + (1-2\mu)e_j^n + \left(\mu - \frac{b\mu\Delta x}{2}\right)e_{j+1}^n + c\Delta t(\Delta t + (\Delta x)^2).$$

Then

$$\left|e_j^{n+1}\right| \le \left|\mu + \frac{b\mu\Delta x}{2}\right|\left|e_{j-1}^n\right| + \left|(1-2\mu)\right|\left|e_j^n\right| + \left|\mu - \frac{b\mu\Delta x}{2}\right|\left|e_{j+1}^n\right| + c\Delta t(\Delta t + (\Delta x)^2).$$

Therefore

$$\left\|e_j^{n+1}\right\|_\infty \le \left|\mu + \frac{b\mu\Delta x}{2}\right|\left\|e_{j-1}^n\right\|_\infty + \left|(1-2\mu)\right|\left\|e_j^n\right\|_\infty + \left|\mu - \frac{b\mu\Delta x}{2}\right|\left\|e_{j+1}^n\right\|_\infty + c\Delta t(\Delta t + (\Delta x)^2).$$

$$\left\|e^{n+1}\right\|_\infty \le \left|\mu + \frac{b\mu\Delta x}{2}\right|\|e^n\|_\infty + \left|(1-2\mu)\right|\|e^n\|_\infty + \left|\mu - \frac{b\mu\Delta x}{2}\right|\|e^n\|_\infty + c\Delta t(\Delta t + (\Delta x)^2).$$

If $1-2\mu \ge 0$ and $\mu - \frac{b\mu\Delta x}{2} \ge 0$, i.e. $\mu \le \frac{1}{2}$ and $1 - \frac{1}{2}b\Delta x > 0$, then

$$\begin{aligned}\left\|e^{n+1}\right\|_\infty &\le \left(\mu + \frac{b\mu\Delta x}{2}\right)\|e^n\|_\infty + ((1-2\mu))\|e^n\|_\infty + \left(\mu - \frac{b\mu\Delta x}{2}\right)\|e^n\|_\infty + c\Delta t(\Delta t + (\Delta x)^2)\\ &= \|e^n\|_\infty + c\Delta t(\Delta t + (\Delta x)^2).\end{aligned}$$

Then

$$\begin{aligned}\|e^n\|_\infty &\le \left\|e^{n-1}\right\|_\infty + c\Delta t(\Delta t + (\Delta x)^2)\\ &\le \left\|e^{n-2}\right\|_\infty + c2\Delta t(\Delta t + (\Delta x)^2)\\ &\le \qquad\vdots\\ &\le \left\|e^0\right\|_\infty + cn\Delta t(\Delta t + (\Delta x)^2)\\ &= ct^*(\Delta t + (\Delta x)^2).\end{aligned}$$

◀

Problem 9.3. *(Prelim Aug. 2010#8) Consider the Crank-Nicolson scheme*

$$u_j^{n+1} = u_j^n + \frac{\mu}{2}\left(u_{j-1}^{n+1} - 2u_j^{n+1} + u_{j+1}^{n+1} + u_{j-1}^n - 2u_j^n + u_{j+1}^n\right)$$

for approximating the solution to the heat equation $\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$ on the intervals $0 \le x \le 1$ and $0 \le t \le t^*$ with the boundary conditions $u(0,t) = u(1,t) = 0$.

1. *Show that the scheme may be written in the form $\mathbf{u}^{n+1} = A\mathbf{u}^n$, where $A \in \mathbb{R}^{m\times m}_{sym}$(the space of $m \times m$ symmetric matrices) and*

$$\|Ax\|_2 \le \|x\|_2,$$

for any $\mathbf{x} \in \mathbb{R}^m$, regardless of the value of μ.

2. *Show that*

$$\|Ax\|_\infty \le \|x\|_\infty,$$

for any $\mathbf{x} \in \mathbb{R}^m$, provided $\mu \le 1$.(In other words, the scheme may only be conditionally stable in the max norm.)

Solution. 1. the scheme

$$u_j^{n+1} = u_j^n + \frac{\mu}{2}\left(u_{j-1}^{n+1} - 2u_j^{n+1} + u_{j+1}^{n+1} + u_{j-1}^n - 2u_j^n + u_{j+1}^n\right)$$

can be rewritten as

$$-\frac{\mu}{2}u_{j-1}^{n+1} + (1+\mu)u_j^{n+1} - \frac{\mu}{2}u_{j+1}^{n+1} = \frac{\mu}{2}u_{j-1}^n + (1-\mu)u_j^n + \frac{\mu}{2}u_{j+1}^n.$$

By using the boundary, we have

$$C\mathbf{u}^{n+1} = B\mathbf{u}^n$$

where

$$C = \begin{bmatrix} 1+\mu & -\frac{\mu}{2} & & & \\ -\frac{\mu}{2} & 1+\mu & -\frac{\mu}{2} & & \\ & \ddots & \ddots & \ddots & \\ & & -\frac{\mu}{2} & 1+\mu & -\frac{\mu}{2} \\ & & & -\frac{\mu}{2} & 1+\mu \end{bmatrix}, B = \begin{bmatrix} 1-\mu & \frac{\mu}{2} & & & \\ \frac{\mu}{2} & 1-\mu & \frac{\mu}{2} & & \\ & \ddots & \ddots & \ddots & \\ & & \frac{\mu}{2} & 1-\mu & \frac{\mu}{2} \\ & & & \frac{\mu}{2} & 1-\mu \end{bmatrix},$$

$$\mathbf{u}^{n+1} = \begin{bmatrix} u_1^{n+1} \\ u_2^{n+1} \\ \vdots \\ u_m^{n+1} \end{bmatrix} \text{ and } \mathbf{u}^n = \begin{bmatrix} u_1^n \\ u_2^n \\ \vdots \\ u_m^n \end{bmatrix}.$$

So, the scheme may be written in the form $\mathbf{u}^{n+1} = A\mathbf{u}^n$, where $A = C^{-1}B$. By using the Fourier transform, i.e.

$$u_j^{n+1} = g(\xi)u_j^n, \;\; u_j^n = e^{ij\Delta x\xi},$$

then we have

$$-\frac{\mu}{2}g(\xi)u_{j-1}^n + (1+\mu)g(\xi)u_j^n - \frac{\mu}{2}g(\xi)u_{j+1}^n = \frac{\mu}{2}u_{j-1}^n + (1-\mu)u_j^n + \frac{\mu}{2}u_{j+1}^n.$$

And then

$$-\frac{\mu}{2}g(\xi)e^{i(j-1)\Delta x\xi} + (1+\mu)g(\xi)e^{ij\Delta x\xi} - \frac{\mu}{2}g(\xi)e^{i(j+1)\Delta x\xi} = \frac{\mu}{2}e^{i(j-1)\Delta x\xi} + (1-\mu)e^{ij\Delta x\xi} + \frac{\mu}{2}e^{i(j+1)\Delta x\xi},$$

i.e.

$$g(\xi)\left(-\frac{\mu}{2}e^{-i\Delta x\xi} + (1+\mu) - \frac{\mu}{2}e^{i\Delta x\xi}\right)e^{j\Delta x\xi} = \left(\frac{\mu}{2}e^{-i\Delta x\xi} + (1-\mu) + \frac{\mu}{2}e^{i\Delta x\xi}\right)e^{j\Delta x\xi},$$

i.e.

$$g(\xi)\left(1+\mu-\mu\cos(\Delta x\xi)\right) = 1-\mu+\mu\cos(\Delta x\xi).$$

therefore,

$$g(\xi) = \frac{1-\mu+\mu\cos(\Delta x\xi)}{1+\mu-\mu\cos(\Delta x\xi)}.$$

hence

$$g(\xi) = \frac{1+\frac{1}{2}z}{1-\frac{1}{2}z}, z = 2\frac{\Delta t}{\Delta x^2}(cos(\Delta x\xi)-1).$$

Moreover, $\left|g(\xi)\right| < 1$, therefore, $\rho(A) < 1$.

$$\|Ax\|_2 \le \|A\|_2 \|x\|_2 = \rho(A)\|x\|_2 \le \|x\|_2.$$

2. the scheme

$$u_j^{n+1} = u_j^n + \frac{\mu}{2}\left(u_{j-1}^{n+1} - 2u_j^{n+1} + u_{j+1}^{n+1} + u_{j-1}^n - 2u_j^n + u_{j+1}^n\right)$$

can be rewritten as

$$(1+\mu)u_j^{n+1} = \frac{\mu}{2}u_{j-1}^{n+1} + \frac{\mu}{2}u_{j+1}^{n+1} + \frac{\mu}{2}u_{j-1}^n + (1-\mu)u_j^n + \frac{\mu}{2}u_{j+1}^n.$$

then

$$\left|1+\mu\right|\left|u_j^{n+1}\right| \le \left|\frac{\mu}{2}\right|\left|u_{j-1}^{n+1}\right| + \left|\frac{\mu}{2}\right|\left|u_{j+1}^{n+1}\right| + \left|\frac{\mu}{2}\right|\left|u_{j-1}^n\right| + \left|(1-\mu)\right|\left|u_j^n\right| + \left|\frac{\mu}{2}\right|\left|u_{j+1}^n\right|.$$

Therefore

$$(1+\mu)\left\|u_j^{n+1}\right\|_\infty \le \frac{\mu}{2}\left\|u_{j-1}^{n+1}\right\|_\infty + \frac{\mu}{2}\left\|u_{j+1}^{n+1}\right\|_\infty + \frac{\mu}{2}\left\|u_{j-1}^n\right\|_\infty + \left|(1-\mu)\right|\left\|u_j^n\right\|_\infty + \frac{\mu}{2}\left\|u_{j+1}^n\right\|_\infty.$$

i.e.

$$(1+\mu)\left\|\mathbf{u}^{n+1}\right\|_\infty \le \frac{\mu}{2}\left\|\mathbf{u}^{n+1}\right\|_\infty + \frac{\mu}{2}\left\|\mathbf{u}^{n+1}\right\|_\infty + \frac{\mu}{2}\|\mathbf{u}^n\|_\infty + \left|(1-\mu)\right|\|\mathbf{u}^n\|_\infty + \frac{\mu}{2}\|\mathbf{u}^n\|_\infty.$$

if $\mu \le 1$, then

$$\left\|\mathbf{u}^{n+1}\right\|_\infty \le \|\mathbf{u}^n\|_\infty,$$

i.e.

$$\|A\mathbf{u}^n\|_\infty \le \|\mathbf{u}^n\|_\infty.$$

◀

Problem 9.4. *(Prelim Aug. 2010#9) Consider the Lax-Wendroff scheme*

$$u_j^{n+1} = u_j^n + \frac{a^2(\Delta t)^2}{2(\Delta x)^2}\left(u_{j-1}^n - 2u_j^n + u_{j+1}^n\right) - \frac{a\Delta t}{2\Delta x}\left(u_{j+1}^n - u_{j-1}^n\right),$$

for the approximating the solution of the Cauchy problem for the advection equation

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0, a > 0.$$

Use Von Neumann's Method to show that the Lax-Wendroff scheme is stable provided the CFL condition

$$\frac{a\Delta t}{\Delta x} \le 1.$$

is enforced.

Solution. By using the Fourier transform, i.e.

$$u_j^{n+1} = g(\xi)u_j^n, \quad u_j^n = e^{ij\Delta x\xi},$$

then we have

$$g(\xi)u_j^n = u_j^n + \frac{a^2(\Delta t)^2}{2(\Delta x)^2}\left(u_{j-1}^n - 2u_j^n + u_{j+1}^n\right) - \frac{a\Delta t}{2\Delta x}\left(u_{j+1}^n - u_{j-1}^n\right).$$

And then

$$g(\xi)e^{ij\Delta x\xi} = e^{ij\Delta x\xi} + \frac{a^2(\Delta t)^2}{2(\Delta x)^2}\left(e^{i(j-1)\Delta x\xi} - 2e^{ij\Delta x\xi} + e^{i(j+1)\Delta x\xi}\right) - \frac{a\Delta t}{2\Delta x}\left(e^{i(j+1)\Delta x\xi} - e^{i(j-1)\Delta x\xi}\right).$$

Therefore

$$\begin{aligned} g(\xi) &= 1 + \frac{a^2(\Delta t)^2}{2(\Delta x)^2}\left(e^{-i\Delta x\xi} - 2 + e^{i\Delta x\xi}\right) - \frac{a\Delta t}{2\Delta x}\left(e^{i\Delta x\xi} - e^{-i\Delta x\xi}\right) \\ &= 1 + \frac{a^2(\Delta t)^2}{2(\Delta x)^2}\left(2\cos(\Delta x\xi) - 2\right) - \frac{a\Delta t}{2\Delta x}\left(2i\sin(\Delta x\xi)\right) \\ &= 1 + \frac{a^2(\Delta t)^2}{(\Delta x)^2}\left(\cos(\Delta x\xi) - 1\right) - \frac{a\Delta t}{\Delta x}\left(i\sin(\Delta x\xi)\right). \end{aligned}$$

Let $\mu = \frac{a\Delta t}{\Delta x}$, then

$$g(\xi) = 1 + \mu^2\left(\cos(\Delta x\xi) - 1\right) - \mu\left(i\sin(\Delta x\xi)\right).$$

If $\left|g(\xi)\right| < 1$, then the scheme is stable, i,e

$$\left(1 + \mu^2\left(\cos(\Delta x\xi) - 1\right)\right)^2 + \left(\mu\sin(\Delta x\xi)\right)^2 < 1.$$

i.e.

$$1 + 2\mu^2\left(\cos(\Delta x\xi) - 1\right) + \mu^4\left(\cos(\Delta x\xi) - 1\right)^2 + \mu^2\sin(\Delta x\xi)^2 < 1.$$

i.e.

$$\mu^2\left(\sin(\Delta x\xi)^2 + 2\cos(\Delta x\xi) - 2\right) + \mu^4\left(\cos(\Delta x\xi) - 1\right)^2 < 0.$$

i.e.

$$\mu^2\left(1 - \cos(\Delta x\xi)^2 + 2\cos(\Delta x\xi) - 2\right) + \mu^4\left(\cos(\Delta x\xi) - 1\right)^2 < 0.$$

i.e

$$\mu^2(\cos(\Delta x\xi) - 1)^2 - (\cos(\Delta x\xi) - 1)^2 < 0,$$

$$(\mu^2 - 1)(\cos(\Delta x\xi) - 1)^2 < 0,$$

then we get $\mu < 1$. The above process is invertible, therefore, we prove the result. ◀

Problem 9.5. *(Prelim Aug. 2010#9)*

Solution. ◀

10 Finite Element Method

Theorem 10.1. *(1D Dirichlet-Poincaré inequality) Let $a > 0$, $u \in C^1([-a,a])$ and $u(-a) = 0$, then the 1D Dirichlet-Poincaré inequality is defined as follows*

$$\int_{-a}^{a} |u(x)|^2 \, dx \leq 4a^2 \int_{-a}^{a} |u'(x)|^2 \, dx.$$

Proof. Since $u(-a) = 0$, then by calculus fact, we have

$$u(x) = u(x) - u(-a) = \int_{-a}^{x} u'(\xi) d\xi.$$

Therefore

$$\begin{aligned}
|u(x)| &\leq \left| \int_{-a}^{x} u'(\xi) d\xi \right| \\
&\leq \int_{-a}^{x} |u'(\xi)| d\xi \\
&\leq \int_{-a}^{a} |u'(\xi)| d\xi \ (x \leq a) \\
&\leq \left(\int_{-a}^{a} 1^2 d\xi \right)^{1/2} \left(\int_{-a}^{a} |u'(\xi)|^2 d\xi \right)^{1/2} \text{ (Cauchy-Schwarz inequality)} \\
&= (2a)^{1/2} \left(\int_{-a}^{a} |u'(\xi)|^2 d\xi \right)^{1/2}.
\end{aligned}$$

Therefore

$$|u(x)|^2 \leq 2a \int_{-a}^{a} |u'(\xi)|^2 d\xi.$$

Integration on both sides of the above equation from $-a$ to a w.r.t. x yields

$$\begin{aligned}
\int_{-a}^{a} |u(x)|^2 \, dx &\leq \int_{-a}^{a} 2a \int_{-a}^{a} |u'(\xi)|^2 \, d\xi dx \\
&= \int_{-a}^{a} |u'(\xi)|^2 \, d\xi \int_{-a}^{a} 2a dx \\
&= 4a^2 \int_{-a}^{a} |u'(\xi)|^2 \, d\xi \\
&= 4a^2 \int_{-a}^{a} |u'(x)|^2 \, dx.
\end{aligned}$$

□

Theorem 10.2. *(1D Neumann-Poincaré inequality) Let $a > 0$, $u \in C^1([-a,a])$ and $\bar{u} = \fint_{-a}^{a} u(x)dx$, then the 1D Neumann-Poincaré inequality is defined as follows*

$$\int_{-a}^{a} |u(x) - \bar{u}(x)|^2 \, dx \leq 2a(a-c) \int_{-a}^{a} |u'(x)|^2 \, dx.$$

Proof. Since, $\bar{u} = \fint_{-a}^{a} u(x)dx$, then by intermediate value theorem, there exists a $c \in [-a,a]$, s.t.

$$u(c) = \bar{u}(x).$$

then by calculus fact, we have

$$u(x) - \bar{u}(x) = u(x) - u(c) = \int_c^x u'(\xi)d\xi.$$

Therefore

$$\begin{aligned}
\left|u(x) - \bar{u}(x)\right| &\leq \left|\int_c^x u'(\xi)d\xi\right| \\
&\leq \int_c^x \left|u'(\xi)\right| d\xi \\
&\leq \int_c^a \left|u'(\xi)\right| d\xi \ (x \leq a) \\
&\leq \left(\int_c^a 1^2 d\xi\right)^{1/2} \left(\int_c^a \left|u'(\xi)\right|^2 d\xi\right)^{1/2} \ \text{(Cauchy-Schwarz inequality)} \\
&= (a-c)^{1/2} \left(\int_{-a}^a \left|u'(\xi)\right|^2 d\xi\right)^{1/2}.
\end{aligned}$$

Therefore

$$\left|u(x) - \bar{u}(x)\right|^2 \leq (a-c) \int_{-a}^a \left|u'(\xi)\right|^2 d\xi.$$

Integration on both sides of the above equation from $-a$ to a w.r.t. x yields

$$\begin{aligned}
\int_{-a}^a \left|u(x) - \bar{u}(x)\right|^2 dx &\leq \int_{-a}^a (a-c) \int_{-a}^a \left|u'(\xi)\right|^2 d\xi dx \\
&= \int_{-a}^a \left|u'(\xi)\right|^2 d\xi \int_{-a}^a (a-c)dx \\
&= 2a(a-c) \int_{-a}^a \left|u'(\xi)\right|^2 d\xi \\
&= 2a(a-c) \int_{-a}^a \left|u'(x)\right|^2 dx.
\end{aligned}$$

□

Definition 10.1. *(symmetric, continuous and coercive) We consider the bilinear form $a : H \times H \to \mathbb{R}$ on a normed space H.*

1. *$a(\cdot,\cdot)$ is said to be symmetric provided that*

$$a(u,v) = a(v,u),\ \forall u,v \in H.$$

2. *$a(\cdot,\cdot)$ is said to continuous or bounded , if there exists a constant C s.t.*

$$\left|a(u,v)\right| = C\|u\|\|v\|,\ \forall u,v \in H.$$

3. *$a(\cdot,\cdot)$ is said to be coercive provided there exists a constant α s.t.*

$$\left|a(u,u)\right| \geq \alpha\|u\|^2,\ \forall u \in H.$$

Proof. □

Theorem 10.3. *(Lax-Milgram Theorem[1])* *Given a Hilbert space H, a continuous, coercive bilinear form $a(\cdot,\cdot)$ and a continuous functional $F \in H'$, there exists a unique $u \in H$ s.t.*

$$a(u,v) = F(v),\ \forall v \in H.$$

Theorem 10.4. *(Céa Lemma[1])* *Suppose V is subspace of H. $a(\cdot,\cdot)$ is continuous and coercive bilinear form on V. Given $F \in V'$, $u \in V$, s.t.*

$$a(u,v) = F(v),\ \forall v \in V.$$

For the finite element variational problem

$$a(u_h,v) = F(v),\ \forall v \in V_h,$$

we have

$$\|u-u_h\|_V \le \frac{C}{\alpha}\min_{v\in V_h}\|u-v\|_V,$$

where C is the continuity constant and α is the coercivity constant of $a(\cdot,\cdot)$ on V.

Proof. □

10.1 Finite element methods for 1D elliptic problems

Theorem 10.5. *(Convergence of 1d FEM)* *The linear basis FEM solution u_h for*

$$\begin{cases} -u''(x) = f(x),\ x \in I = \{x \in [a,b]\}, \\ u(a) = u(b) = 0. \end{cases}$$

has the following properties

$$\|u-u_h\|_{L^2(I)} \le Ch^2 \left\|u''\right\|_{L^2(I)},$$
$$\left\|u'-u_h'\right\|_{L^2(I)} \le Ch \left\|u''\right\|_{L^2(I)}.$$

Proof. 1. Define the first degree Taylor polynomial on $I_i = [x_i, x_{i+1}]$ as

$$Q_1 u(x) = u(x_i) + u'(x)(x-x_i).$$

Then we have

$$\left|u(x) - Q_1 u(x)\right| = \int_I \left|x-y\right| u''(y)dy.$$

This implies

$$\begin{aligned}\left\|u(x)-Q_1u(x)\right\|_{C(I_i)} &= \max_{x\in I_i}\int_{I_i}\left|x-y\right|\left|u''(y)\right|dy\\ &\le h\int_{I_i}\left|u''(y)\right|dy\\ &\le h\left(\int_{I_i}1^2dy\right)^{1/2}\left(\int_{I_i}\left|u''(y)\right|^2dy\right)^{1/2}\\ &\le h^{3/2}\left(\int_{I_i}\left|u''(y)\right|^2dy\right)^{1/2}\\ &= h^{3/2}\left\|u''(x)\right\|_{L^2(I_i)}.\end{aligned}$$

And

$$\begin{aligned}\|u-u_h\|^2_{L^2(I_i)} &= \int_{I_i}(u-u_h)^2dx\\ &\le \|u-u_h\|^2_{C(I_i)}\int_{I_i}dx\\ &\le h\|u-u_h\|^2_{C(I_i)}.\end{aligned}$$

Therefore,

$$\|u-u_h\|_{L^2(I_i)}\le h^{1/2}\|u-u_h\|_{C(I_i)}.$$

and

$$\begin{aligned}\|u-u_h\|_{C(I_i)} &\le \|u-Q_1u\|_{C(I_i)}+\|Q_1u-u_h\|_{C(I_i)}\\ &= \|u-Q_1u\|_{C(I_i)}+\left\|I_h(Q_1u-u)\right\|_{C(I_i)}\\ &\le 2\|u-Q_1u\|_{C(I_i)}\\ &= 2h^{3/2}\left\|u''(x)\right\|_{L^2(I_i)}\end{aligned}$$

Therefore

$$\|u-u_h\|_{L^2(I_i)}\le 2h^2\left\|u''(x)\right\|_{L^2(I_i)}.$$

Hence

$$\|u-u_h\|_{L^2(I)}\le 2h^2\left\|u''(x)\right\|_{L^2(I)}.$$

2. For the linear basis we have the Element solution $u_h(x_i)=u(x_i)$ and $u_h(x_{i+1})=u(x_{i+1})$ on element $I_i=[x_i,x_{i+1}]$. and

$$\begin{aligned}u_h'(x) &= \frac{u_h(x_{i+1})-u_h(x_i)}{h}=\frac{u(x_{i+1})-u(x_i)}{h}=\frac{1}{h}\int_{x_i}^{x_{i+1}}u'(y)dy=\frac{1}{h}\int_{I_i}u'(y)dy,\\ u'(x) &= u'(x)\frac{h}{h}=\frac{1}{h}\int_{x_i}^{x_{i+1}}u'(x)dy=\frac{1}{h}\int_{I_i}u'(x)dy.\end{aligned}$$

Therefore

$$\begin{aligned} u_h'(x) - u'(x) &= \frac{1}{h}\int_{I_i} u'(y) - u'(x)dy \\ &= \frac{1}{h}\int_{I_i}\int_x^y u''(\xi)d\xi dy. \end{aligned}$$

so

$$\begin{aligned} \left\|u' - u_h'\right\|^2_{L^2(I_i)} &= \int_{I_i}\left(u_h'(x) - u'(x)\right)^2 dx \\ &= \frac{1}{h^2}\int_{I_i}\left(\int_{I_i}\int_x^y u''(\xi)d\xi dy\right)^2 dx \\ &\le \frac{1}{h^2}\int_{I_i}\left(\int_{I_i}\int_{I_i} u''(\xi)d\xi dy\right)^2 dx \\ &= \frac{1}{h^2}\left(\int_{I_i}\int_{I_i} u''(\xi)d\xi dy\right)^2 \int_{I_i} dx \\ &= \frac{1}{h}\left(\int_{I_i} dy \int_{I_i} u''(\xi)d\xi\right)^2 \\ &= \frac{1}{h}\left(h\int_{I_i} u''(\xi)d\xi\right)^2 \\ &= h\left(\int_{I_i} u''(\xi)d\xi\right)^2 \\ &\le h\left(\left(\int_{I_i} 1^2 d\xi\right)^{1/2}\left(\int_{I_i}\left|u''(\xi)\right|^2 d\xi\right)^{1/2}\right)^2 \\ &\le h^2\left(\int_{I_i}\left|u''(\xi)\right|^2 d\xi\right) \end{aligned}$$

hence

$$\left\|u' - u_h'\right\|_{L^2(I_i)} \le Ch\left\|u''\right\|_{L^2(I_i)}.$$

Therefore,

$$\left\|u' - u_h'\right\|_{L^2(I)} \le Ch\left\|u''\right\|_{L^2(I)}.$$

□

10.2 Problems

Problem 10.1. *(Prelim Jan. 2008#8) Let $\Omega \subset \mathbb{R}^2$ be a bounded domain with a smooth boundary. Consider a 2-D poisson-like equation*

$$\begin{cases} -\Delta u + 3u &= x^2y^2, \text{ in } \Omega, \\ u &= 0, \text{ on } \partial\Omega. \end{cases}$$

1. *Write the corresponding Ritz and Galerkin variational problems.*
2. *Prove that the Galerkin method has a unique solution u_h and the following estimate is valid*

$$\|u - u_h\|_{H^1} \leq C \inf_{v_h \in V_h} \|u - v_h\|_{H^1},$$

with C independent of h, where V_h denotes a finite element subspace of $H^1(\Omega)$ consisting of continuous piecewise polynomials of degree $k \geq 1$.

Solution. 1. For this pure Dirichlet Problem, the test functional space $v \in H_0^1$. Multiple the test function on the both sides of the original function and integral on Ω, we get

$$-\int_\Omega \Delta u v dx + \int_\Omega uv dx = \int_\Omega xyv dx.$$

Integration by part yields

$$\int_\Omega \nabla u \nabla v dx + \int_\Omega uv dx = \int_\Omega xyv dx.$$

Let

$$a(u,v) = \int_\Omega \nabla u \nabla v dx + \int_\Omega uv dx, \quad f(v) = \int_\Omega xyv dx.$$

Then, the

(a) Ritz variational problem is: find $u_h \in H_0^1$, such that

$$J(u_h) = \min \frac{1}{2} a(u_h, u_h) - f(u_h).$$

(b) Galerkin variational problem is: find $u_h \in H_0^1$, such that

$$a(u_h, u_h) = f(u_h).$$

2. Next, we will use Lax-Milgram to prove the uniqueness.

(a)

$$\begin{aligned} \left|a(u,v)\right| &\leq \int_\Omega |\nabla u \nabla v| dx + \int_\Omega |uv| dx \\ &\leq \|\nabla u\|_{L^2(\Omega)} \|\nabla v\|_{L^2(\Omega)} + \|u\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} \\ &\leq \|\nabla u\|_{L^2(\Omega)} \|\nabla v\|_{L^2(\Omega)} + C \|\nabla u\|_{L^2(\Omega)} \|\nabla v\|_{L^2(\Omega)} \\ &\leq C \|\nabla u\|_{L^2(\Omega)} \|\nabla v\|_{L^2(\Omega)} \\ &\leq C \|u\|_{H^1(\Omega)} \|v\|_{H^1(\Omega)} \end{aligned}$$

(b)

$$a(u,u) = \int_\Omega (\nabla u)^2 dx + \int_\Omega u^2 dx$$

So,

$$\begin{aligned} \left|a(u,u)\right| &= \int_\Omega |\nabla u|^2 \, dx + \int_\Omega |u|^2 \, dx \\ &= \|\nabla u\|^2_{L^2(\Omega)} + \|u\|^2_{L^2(\Omega)} \\ &= \|u\|^2_{H^1(\Omega)}. \end{aligned}$$

(c)

$$\begin{aligned} \left|f(v)\right| &\le \int_\Omega \left|xyv\right| dx \\ &\le \max|xy| \int_\Omega |v| \, dx \\ &\le C\left(\int_\Omega 1^2 dx\right)^{1/2} \left(\int_\Omega |v|^2 \, dx\right)^{1/2} \\ &\le C\|v\|_{L^2(\Omega)} \le C\|v\|_{H^1(\Omega)}. \end{aligned}$$

by Lax-Milgram theorem, we get that e Galerkin method has a unique solution u_h. Moreover,

$$a(v_h, v_h) = f(v_h).$$

And from the weak formula, we have

$$a(u, v_h) = f(v_h).$$

then we get the Galerkin Orthogonal (GO)

$$a(u - u_h, v_h) = 0.$$

Then by coercivity

$$\begin{aligned} \|u - u_h\|^2_{H^1(\Omega)} &\le \left|a(u - u_h, u - u_h)\right| \\ &= \left|a(u - u_h, u - v_h) + a(u - u_h, v_h - u_h)\right| \\ &= \left|a(u - u_h, u - v_h)\right| \\ &\le \|u - u_h\|_{H^1(\Omega)} \|u - v_h\|_{H^1(\Omega)}. \end{aligned}$$

Therefore,

$$\|u - u_h\|_{H^1} \le C \inf_{v_h \in V_h} \|u - v_h\|_{H^1},$$

◀

Problem 10.2. *(Prelim Aug. 2006#9) Let* $\Omega := \left\{(x,y) : x^2+y^2<1\right\}$*, consider the poisson problem*

$$\begin{cases} -\Delta u + 2u &= xy, \text{ in } \Omega, \\ u &= 0, \text{ on } \partial\Omega. \end{cases}$$

1. *Define the corresponding Ritz and Galerkin variational formulations.*
2. *Suppose that the Galerkin variational problem has solution, prove that the Ritz variation problem must also have a solution. Is the converse statement true?*
3. *Let* V_N *be an N-dimension subspace of* $W^{1,2}(\Omega)$*. Define the Galerkin method for approximating the solution of the poisson equation problem, and prove that the Galerkin method has a unique solution.*
4. *Let* u_N *denote the Galerkin solution, prove that*

$$\|u-u_N\|_E \le C \inf_{v_N \in V_N} \|u-v_N\|_E,$$

where

$$\|v\|_E := \int_\Omega \left(|\nabla v|^2 + 2v^2\right) dxdy.$$

Solution. ◂

References

[1] S. C. Brenner and R. Scott, *The mathematical theory of finite element methods*, vol. 15, Springer, 2008. 108

[2] A. Iserles, *A First Course in the Numerical Analysis of Differential Equations (Cambridge Texts in Applied Mathematics)*, Cambridge University Press, 2008. 80, 84, 86, 92, 151

[3] Y. Saad, *Iterative methods for sparse linear systems*, Siam, 2003. 2, 39

[4] A. J. Salgado, *Numerical math lecture notes: 571-572.* UTK, 2013-14. 1

[5] S. M. Wise, *Numerical math lecture notes: 571-572.* UTK, 2012-13. 1

Appendices

A Numerical Mathematics Preliminary Examination Sample Question, Summer, 2013

A.1 Numerical Linear Algebra

Problem A.1. *(Sample#1) Suppose $A \in \mathbb{C}^{n\times n}_{her}$, and $\rho(A) \subset (0,\infty)$. Prove that A is Hermitian Positive Definite.*

Solution. Since $A \in \mathbb{C}^{n\times n}_{her}$, then the eigenvalue of A are real. Let λ be arbitrary eigenvalue of A, then

$$(Ax,x) = (\lambda x,x) = \lambda(x,x),$$
$$(Ax,x) = (x,A^*x) = (x,Ax)(x,\lambda x) = \overline{\lambda}(x,x),$$

and then $\lambda = \overline{\lambda}$, so λ is real. Moreover, since $\rho(A) \subset (0,\infty)$, then we have λ is positive.

$$x^*Ax = x^*\lambda x = \lambda x^*x = \lambda(x_1^2 + x_2^2 + \cdots + x_n^2) > 0.$$

for all $x \neq 0$. Hence, A is Hermitian Positive Definite. ◂

Problem A.2. *(Sample#2) Suppose $dim(A) = n$. If A has n distinct eigenvalues , then A is diagonalizable .*

Solution. (Sketch) Suppose $n = 2$, and let λ_1, λ_2 be distinct eigenvalues of A with corresponding eigenvectors v_1, v_2. Now, we will use contradiction to show v_1, v_2 are lineally independent. Suppose v_1, v_2 are lineally dependent, then

$$c_1v_1 + c_2v_2 = 0, \tag{210}$$

with c_1, c_2 are not both 0. Multiplying A on both sides of (210), then

$$c_1Av_1 + c_2Av_2 = c_1\lambda_1v_1 + c_2\lambda_2v_2 = 0. \tag{211}$$

Multiplying λ_1 on both sides of (210), then

$$c_1\lambda_1v_1 + c_2\lambda_1v_2 = 0. \tag{212}$$

Subtracting (212) form (211), then

$$c_2(\lambda_2 - \lambda_1)v_2 = 0. \tag{213}$$

Since $\lambda_1 \neq \lambda_2$ and $v_2 \neq 0$, then $c_2 = 0$. Similarly, we can get $c_1 = 0$. Hence, we get the contradiction.
A similar argument gives the result for n. Then we get A has n linearly independent eigenvectors . ◂

Problem A.3. *(Sample#5) Let $u, v \in \mathbb{C}^n$ and set $A := I_n + uv^* \in \mathbb{C}^{n\times n}$.*

1. *Suppose A is invertible. Prove that $A^{-1} = I_n + \alpha uv^*$, for some $\alpha \in \mathbb{C}$. Give the expression for α.*
2. *For what u and v is A singular ?*
3. *Suppose A is singular. What is the null space of A, N(A), in this case?*

Solution. 1. If $uv^* = 0$, then the proof is trivial. Assume $uv^* \neq 0$, then

$$\begin{aligned} A^{-1}A &= (I_n + \alpha uv^*)(I_n + uv^*) \\ &= I_n + uv^* + \alpha(uv^* + u(v^*u)v^*) \\ &= I_n + (1 + \alpha + \alpha v^*u)uv^* \\ &= I_n. \end{aligned}$$

i.e.

$$1 + \alpha + \alpha v^*u = 0,$$

i.e.

$$\alpha = -\frac{1}{1 + v^*u}, 1 \neq -v^*u.$$

2. For $1 = -v^*u$, the A is singular.
3. If A is singular, then $v^*u = -1$.

Claim A.1. *N(A)=span(u).*

Proof. (a) $\subseteq$ let $w \in N(A)$, then we have

$$Aw = (I_n + uv^*)w = w + uv^*w = 0$$

Then we have $w = -v^*wu$, hence $w \in span(u)$.

(b) $\supseteq$ Let $w \in span(u)$, then we have $w = \beta u$, then

$$Aw = (I_n + uv^*)\beta u = \beta(u + uv^*u) = \beta(u + (v^*u)u) = 0.$$

hence $span(u) \in w$.

◀

◀

Problem A.4. *(Sample #6) Suppose that $A \in \mathbb{R}^{n\times n}$ is SPD.*

1. *Show that $\|x\|_A = \sqrt{x^TAx}$ defines a vector norm.*
2. *Let the eigenvalues of A be ordered so that $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. Show that*

$$\sqrt{\lambda_1}\|x\|_2 \leq \|x\|_A \leq \sqrt{\lambda_n}\|x\|_2.$$

for any $x \in \mathbb{R}^n$.

3. *Let $b \in \mathbb{R}^n$ be given. Prove that $x_* \in \mathbb{R}^n$ solves $Ax = b$ if and only if x_* minimizes the quadratic function $f : \mathbb{R}^n \to \mathbb{R}$ defined by*

$$f(x) = \frac{1}{2}x^TAx - x^Tb.$$

Solution. 1. (a) Obviously, $\|x\|_A = \sqrt{x^TAx} \geq 0$. When $x = 0$, then $\|x\|_A = \sqrt{x^TAx} = 0$; when $\|x\|_A = \sqrt{x^TAx} = 0$, then we have $(Ax, x) = 0$, since A is SPD, therefore, $x \equiv 0$.

(b) $\|\lambda x\|_A = \sqrt{\lambda x^TA\lambda x} = \sqrt{\lambda^2x^TAx} = |\lambda|\sqrt{x^TAx} = |\lambda|\|x\|_A$.

(c) Next we will show $\|x+y\|_A = \|x\|_A + \|y\|_A$. First, we would like to show

$$\left|y^T Ax\right| \le \|x\|_A \|y\|_A .$$

Since A is SPD, therefore $A = R^T R$, moreover

$$\|Rx\|_2 = (Rx, Rx)^{1/2} = \sqrt{(Rx)^T Rx} = \sqrt{x^T R^T Rx} = \sqrt{x^T Ax} = \|x\|_A .$$

Then

$$\left|y^T Ax\right| = \left|y^T R^T Rx\right| = \left|(Ry)^T Rx\right| = \left|(Rx, Ry)\right| \overset{c.s.}{\le} \|Rx\|_2 \|Ry\|_2 = \|x\|_A \|y\|_A .$$

And

$$\begin{aligned} \|x+y\|_A^2 &= (x+y, x+y)_A = (x,x)_A + 2(x,y)_A + (y,y)_A \\ &\le \|x\|_A + 2\left|y^T Ax\right| + \|y\|_A \\ &\le \|x\|_A + 2\|x\|_A \|y\|_A + \|y\|_A \\ &= \left(\|x\|_A + \|y\|_A\right)^2 . \end{aligned}$$

therefore

$$\|x+y\|_A = \|x\|_A + \|y\|_A .$$

2. Since A is SPD, therefore $A = R^T R$, moreover

$$\|Rx\|_2 = (Rx, Rx)^{1/2} = \sqrt{(Rx)^T Rx} = \sqrt{x^T R^T Rx} = \sqrt{x^T Ax} = \|x\|_A .$$

Let $0 < \tilde{\lambda}_1 \le \tilde{\lambda}_2 \le \cdots \le \tilde{\lambda}_n$ be the eigenvalue of R, then $\tilde{\lambda}_i = \sqrt{\lambda_i}$. so

$$\left|\tilde{\lambda}_1\right| \|x\|_2 \le \|Rx\|_2 = \|x\|_A \le \left|\tilde{\lambda}_n\right| \|x\|_2 .$$

i.e.

$$\sqrt{\lambda_1} \|x\|_2 \le \|Rx\|_2 = \|x\|_A \le \sqrt{\lambda_n} \|x\|_2 .$$

3. Since

$$\begin{aligned} \frac{\partial}{\partial x_i}\left(x^T Ax\right) &= \frac{\partial}{\partial x_i}\left(x^T\right) Ax + x^T \frac{\partial}{\partial x_i}(Ax) \\ &= [0, \cdots, 0, \underset{i}{1}, 0, \cdots, 0] Ax + x^T A \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} i \\ &= (Ax)_i + \left(A^T x\right)_i = 2(Ax)_i . \end{aligned}$$

and

$$\frac{\partial}{\partial x_i}\left(x^T b\right) = \frac{\partial}{\partial x_i}\left(x^T\right) b = [0, \cdots, 0, \underset{i}{1}, 0, \cdots, 0] b = b_i .$$

Therefore,

$$\nabla f(x) = \frac{1}{2}2Ax - b = Ax - b.$$

If $Ax_* = b$, then $\nabla f(x_*) = Ax_* - b = 0$, therefore x_* minimizes the quadratic function f. Conversely, when x_* minimizes the quadratic function f, then $\nabla f(x_*) = Ax_* - b = 0$, therefore $Ax_* = b$.

◀

Problem A.5. *(Sample#9) Suppose that the spectrum of $A \in \mathbb{R}^{n\times n}_{sym}$ is denoted $\rho(A) = \{\lambda_1, \lambda_2, \cdots, \lambda_n\} \subset \mathbb{R}$. Let $S = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ be the orthonormal basis of eigenvectors of A, with $A\mathbf{x}_k = \lambda_k \mathbf{x}_k$, for $k = 1, \cdots, n$. The Rayleigh quotient of $x \in \mathbb{R}^n_*$ is defined as*

$$R(\mathbf{x}) := \frac{\mathbf{x}^T A\mathbf{x}}{\mathbf{x}^T\mathbf{x}}.$$

Prove the following facts:

1. $$R(\mathbf{x}) := \frac{\sum_{j=1}^n \lambda_j \alpha_j^2}{\sum_{j=1}^n \alpha_j^2}$$

   where $\alpha_j = \mathbf{x}^T\mathbf{x}_j$.

2. $$\min_{\lambda\in\rho(A)} \lambda \le R(\mathbf{x}) \le \max_{\lambda\in\rho(A)} \lambda.$$

Solution. 1. First, we need to show that $\mathbf{x} = \sum_{j=1}^n \alpha_j x_j$ is the unique representation of $\mathbf{x}$ w.r.t. the orthonormal basis S. Since $S = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ is the orthonormal basis of eigenvectors of A, then $\sum_{j=1}^n \mathbf{x}^T\mathbf{x}_j x_j$ is the representation of $\mathbf{x}$. Assume $\sum_{j=1}^n \beta_j x_j$ is another representation of $\mathbf{x}$. Then we have $\sum_{j=1}^n (\beta_j - \alpha_j)x_j = 0$, since $x_j \neq 0$, so $\alpha = \beta$. Now , we have

$$\begin{aligned}
\mathbf{x}^T A\mathbf{x} &= \mathbf{x}^T A \sum_{j=1}^n \alpha_j \mathbf{x}_j \\
&= \mathbf{x}^T \sum_{j=1}^n \alpha_j A\mathbf{x}_j \\
&= \mathbf{x}^T \sum_{j=1}^n \alpha_j \lambda_j \mathbf{x}_j \\
&= \sum_{j=1}^n \alpha_j \lambda_j \mathbf{x}^T\mathbf{x}_j \\
&= \sum_{j=1}^n \lambda_j \alpha_j^2.
\end{aligned}$$

Similarly, we have $\mathbf{x}^T\mathbf{x} = \sum_{j=1}^n \alpha_j^2$. Hence,

$$R(\mathbf{x}) := \frac{\mathbf{x}^T A\mathbf{x}}{\mathbf{x}^T\mathbf{x}} = \frac{\sum_{j=1}^n \lambda_j \alpha_j^2}{\sum_{j=1}^n \alpha_j^2}.$$

2. Since,

$$R(\mathbf{x}) := \frac{\mathbf{x}^T A\mathbf{x}}{\mathbf{x}^T\mathbf{x}} = \frac{\sum_{j=1}^n \lambda_j \alpha_j^2}{\sum_{j=1}^n \alpha_j^2}.$$

, then

$$\min_j \lambda_j \frac{\sum_{j=1}^n \alpha_j^2}{\sum_{j=1}^n \alpha_j^2} \le R(\mathbf{x}) \le \max_j \lambda_j \frac{\sum_{j=1}^n \alpha_j^2}{\sum_{j=1}^n \alpha_j^2},$$

i.e.

$$\min_j \lambda_j \le R(\mathbf{x}) \le \max_j \lambda_j.$$

Hence

$$\min_{\lambda\in\rho(A)} \lambda \le R(\mathbf{x}) \le \max_{\lambda\in\rho(A)} \lambda.$$

◀

Problem A.6. *(Sample #31) Let $A \in \mathbb{R}^{n\times n}$ be symmetric positive define (SPD). Let $b \in \mathbb{R}^n$. Consider solving $Ax = b$ using the iterative method*

$$Mx^{n+1} = Nx^n + b, \quad n = 0, 1, 2, \cdots$$

where $A = M - N$, M is invertible, and $x^0 \in \mathbb{R}^n$ us arbitrary.

1. *If $M + M^T - A$ is SPD, prove that the method is convergent.*
2. *Prove that the Gauss-Seidel Method converges.*

Solution. 1. From the problem, we get

$$x^{n+1} = M^{-1}Nx^n + M^{-1}b.$$

Let $G = M^{-1}N = M^{-1}(M-A) = I - M^{-1}A$, If we can prove that $\rho(G) < 1$, then this method converges. Let λ be any eigenvalue of G and x be the corresponding eigenvector, i.e.

$$Gx = \lambda x.$$

then

$$(I - M^{-1}A)x = \lambda x,$$

i.e.

$$(M - A)x = \lambda Mx,$$

i.e.

$$(1 - \lambda)Mx = Ax.$$

(a) $\lambda \neq 1$. If $\lambda = 1$, then $Ax = 0$, for any x, so $A = 0$ which contradicts to A is SPD.

(b) $\lambda \leq 1$. Since, $(1-\lambda)Mx = Ax$. then

$$(1-\lambda)x^*Mx = x^*Ax.$$

So we have

$$x^*Mx = \frac{1}{1-\lambda}x^*Ax.$$

taking conjugate transpose of which yields

$$x^*M^*x = \frac{1}{1-\overline{\lambda}}x^*A^*x = \frac{1}{1-\overline{\lambda}}x^*Ax.$$

Then, we have

$$\begin{aligned} x^*(M+M^*-A)x &= \left(\frac{1}{1-\lambda}+\frac{1}{1-\overline{\lambda}}-1\right)x^*Ax \\ &= \left(\frac{\lambda}{1-\lambda}+\frac{1}{1-\overline{\lambda}}\right)x^*Ax \\ &= \frac{1-\lambda^2}{|1-\lambda|^2}x^*Ax. \end{aligned}$$

Since $M+M^*-A$ and A are SPD, then $x^*(M+M^*-A)x > 0$, $x^*Ax > 0$. Therefore,

$$1-\lambda^2 > 0.$$

i.e.

$$|\lambda| < 1.$$

2.

$$\begin{aligned} \text{Jacobi Method:} \quad & M_J = D, N_J = -(L+U) \\ \text{Gauss-Seidel Method:} \quad & M_{GS} = D+L, N_{GS} = -U. \end{aligned}$$

where $A = L+D+U$. Since A is SPD, then $M_{GS}+M_{GS}^T-A = D+L+D^T+L^T-A = D+L^T-U$ is SPD. Therefore, From the part 1, we get that the Gauss-Seidel Method converges.

◀

Problem A.7. *(Sample #32) Let $A \in \mathbb{R}^{n\times n}$ be symmetric positive define (SPD). Let $b \in \mathbb{R}^n$. Consider solving $Ax = b$ using the iterative method*

$$Mx^{n+1} = Nx^n + b, \quad n = 0,1,2,\cdots$$

where $A = M-N$, M is invertible, and $x^0 \in \mathbb{R}^n$ us arbitrary. Suppose that $M+M^T-A$ is SPD. Show that each step of this method reduces the A-norm of $e^n = x-x^n$, whenever $e^n \neq 0$. Recall that, the A-norm of any $y \in \mathbb{R}^n$ is defined via

$$\|y\|_A = \sqrt{y^TAy}.$$

Solution. Let $e^k = x^k - x$. And rewrite the scheme to the canonical form $B = M, \alpha = 1$, then

$$B\left(\frac{x^{k+1}-x^k}{\alpha}\right) + Ax^k = b = Ax.$$

so, we get

$$B\left(\frac{e^{k+1}-e^k}{\alpha}\right) + Ae^k = 0.$$

Let $v^{k+1} = e^{k+1} - e^k$, then

$$\frac{1}{\alpha}Bv^{k+1} + Ae^k = 0.$$

Taking the conjugate transport of the above equation, then we get

$$\frac{1}{\alpha}B^*v^{k+1} + A * e^k = \frac{1}{\alpha}B^*v^{k+1} + Ae^k = 0.$$

therefore

$$\frac{1}{\alpha}(\frac{B+B^*}{2})v^{k+1} + Ae^k = 0.$$

Let $B_s = \frac{B+B^*}{2}$. Then take the inner product of both sides with v^{k+1},

$$\frac{1}{\alpha}(B_s v^{k+1}, v^{k+1}) + (Ae^k, v^{k+1}) = 0.$$

Since

$$e^k = \frac{1}{2}(e^{k+1}+e^k) - \frac{1}{2}(e^{k+1}-e^k) = \frac{1}{2}(e^{k+1}+e^k) - \frac{1}{2}v^{k+1}.$$

Therefore,

$$\begin{aligned} 0 &= \frac{1}{\alpha}(B_s v^{k+1}, v^{k+1}) + (Ae^k, v^{k+1}) \\ &= \frac{1}{\alpha}(B_s v^{k+1}, v^{k+1}) + \frac{1}{2}(A(e^{k+1}+e^k), v^{k+1}) - \frac{1}{2}(Av^{k+1}, v^{k+1}) \\ &= \frac{1}{\alpha}((B_s - \frac{\alpha}{2}A)v^{k+1}, v^{k+1}) + \frac{1}{2}(A(e^{k+1}+e^k), v^{k+1}) \\ &= \frac{1}{\alpha}((B_s - \frac{\alpha}{2}A)v^{k+1}, v^{k+1}) + \frac{1}{2}(\left\|e^{k+1}\right\|_A^2 - \left\|e^k\right\|_A^2) \end{aligned}$$

By assumption, $Q = B_s - \frac{\alpha}{2}A = \frac{M+M^T-A}{2} > 0$, i.e. there exists $m > 0$, s.t.

$$(Qy, y) \geq m\left\|y\right\|_2^2.$$

Therefore,

$$\frac{m}{\alpha}\left\|v^{k+1}\right\|_2^2 + \frac{1}{2}(\left\|e^{k+1}\right\|_A^2 - \left\|e^k\right\|_A^2) \leq 0.$$

i.e.

$$\frac{2m}{\alpha}\left\|v^{k+1}\right\|_2^2 + \left\|e^{k+1}\right\|_A^2 \leq \left\|e^k\right\|_A^2.$$

Hence

$$\left\|e^{k+1}\right\|_A^2 \leq \left\|e^k\right\|_A^2.$$

and

$$\left\|e^{k+1}\right\|_A^2 \to 0.$$

◀

Problem A.8. *(Sample #33) Consider a linear system $Ax = b$ with $A \in \mathbb{R}^{n\times n}$. Richardson's method is an iteration method*

$$Mx^{k+1} = Nx^k + b$$

with $M = \frac{1}{w}I, N = M - A = \frac{1}{w}I - A$, where w is a damping factor chosen to make M approximate A as well as possible. Suppose A is positive definite and $w > 0$. Let λ_1 and λ_n denote the smallest and largest eigenvalues of A.

1. *Prove that Richardson's method converges if only if $w < \frac{2}{\lambda_n}$.*
2. *Prove that the optimal value value of w is $w_0 = \frac{2}{\lambda_1+\lambda_n}$.*

Solution. 1. From the scheme of the Richardson's method, we know that

$$x^{k+1} = (I - wA)x^k + wb.$$

So the error transfer operator is $T = I - wA$. Then if λ_i is the eigenvalue of A, then $1 - w\lambda_i$ should be the eigenvalue of T. The sufficient and necessary condition of convergence is $\rho(T) < 1$, i.e.

$$|1 - w\lambda_i| < 1$$

for all i. Therefore, we have

$$w < \frac{2}{\lambda_i}.$$

Since λ_n denote the largest eigenvalues of A, then $\frac{2}{\lambda_n} \leq \frac{2}{\lambda_i}$. Hence, we need

$$w < \frac{2}{\lambda_n}.$$

conversely, if $w < \frac{2}{\lambda_n}$, then $\rho(T) < 1$, then the scheme converges.

2. The minimum is attachment at $|1 - \omega\lambda_n| = |1 - \omega\lambda_1|$(Figure.1), i.e.

$$\omega\lambda_n - 1 = 1 - \omega\lambda_1.$$

Therefore, we get

$$\omega_{opt} = \frac{2}{\lambda_1 + \lambda_n}.$$

◀

Problem A.9. *(Sample #34) Let $A \in \mathbb{C}^{n\times n}$. Define*

$$S_n := I + A + A^2 + \cdots + A^n.$$

1. *Prove that the sequence $\{S_n\}_{n=0}^{\infty}$ converges if only if A is convergent.*
2. *Prove that if A is convergent, then $I - A$ is non-singular and*

$$\lim_{n\to\infty} S_n = (I-A)^{-1}.$$

Solution. 1. From the problem, we know that

$$S_n := I + A + A^2 + \cdots + A^n = A^0 + A + A^2 + \cdots + A^n = \sum_{k=0}^{n} A^k.$$

Moreover,

$$\left\|A^k\right\| = \sup_{0\neq x\in\mathbb{C}^n} \frac{\left\|A^k x\right\|}{\|x\|} \leq \sup_{0\neq x\in\mathbb{C}^n} \frac{\|A\|\left\|A^{k-1}x\right\|}{\|x\|} \leq \cdots \leq \|A\|^k.$$

From the properties of geometric series, S_n converges if only if $|A| < 1$. Therefore, we get if $|A| < 1$ then A is convergent. Conversely, if A is convergent, then $|A| < 1$. Hence S_n converges.

2.

$$\left\|(I-A)x\right\| = \|x - Ax\| \geq \|x\| - \|Ax\| \geq \|x\| - \|A\|\|x\| = (1 - \|A\|)\|x\|.$$

If A is convergent, then $\|A\| \neq 0$. Therefore, if $\left\|(I-A)x\right\| = 0$, then $\|x\| = 0$, i.e. $ker(I-A) = 0$. Hence, $I - A$ is non-singular. From the definition of S_n, we get

$$(I-A)S_n = \sum_{k=0}^{n} A^k - \sum_{k=1}^{n+1} A^k = A^0 - A^{n+1} = I - A^{n+1}.$$

Taking limit on both sides of the above equation with the fact $|A| < 1$, then we get

$$(I-A)\lim_{n\to\infty} S_n = I.$$

Since $I - A$ is non-singular, then we have

$$\lim_{n\to\infty} S_n = (I-A)^{-1}.$$

◀

Problem A.10. *(Sample #40) Show that if λ is an eigenvalue of A^*A, where $A \in \mathbb{C}^{n\times n}$, then*

$$0 \leq \lambda \leq \|A\|\|A^*\|$$

Solution. Since $x^*A^*Ax = (Ax)^*(Ax) = \lambda x^*x \geq 0$, therefore $\lambda \geq 0$, and λ is real. Since

$$A^*Ax = \lambda x.$$

so

$$0 \leq \lambda\|x\| = \|\lambda x\| = \|A^*Ax\| \leq \|A^*\|\|A\|\|x\|.$$

◀

Problem A.11. *(Sample #41) Suppose $A \in \mathbb{C}^{n\times n}$ and A is invertible. Prove that*

$$\kappa_2 \le \sqrt{\frac{\lambda_n}{\lambda_1}}.$$

where λ_n is the largest eigenvalue of $B := A^*A$, and λ_1 is the smallest eigenvalue of B.

Solution. Since $\kappa_2 = \|A\|_2 \left\|A^{-1}\right\|_2$ and $\|A\|_2 = \max \rho(A) = \sqrt{\lambda_n}$. therefore

$$\kappa_2 = \|A\|_2 \left\|A^{-1}\right\|_2 = \frac{\sqrt{\lambda_n}}{\sqrt{\lambda_1}}.$$

◀

Problem A.12. *(Sample #34) Let $A = [a_{i,j}] \in \mathbb{C}^{n\times n}$ be invertible and $b \in \mathbb{C}^n$. Prove that the classical Jacobi iteration method for approximating the solution to $Ax = b$ is convergent, for any starting value x_0, if A is strictly diagonally dominant, i.e.*

$$\left|a_{i,i}\right| < \sum_{k\neq i} \left|a_{i,k}\right|, \ \forall\ i = 1, \cdots, n.$$

Solution. The Jacobi iteration scheme is as follows

$$D(x^{k+1} - x^k) + Ax^k = b.$$

This scheme can be rewritten as

$$x^{k+1} = (I - D^{-1}A)x^k + D^{-1}b.$$

We want to show If A is diagonal dominant , then $\left\|T_J\right\| < 1$, then Jacobi Method convergences. From the definition of T, we know that T for Jacobi Method is as follows

$$T_J = I - D^{-1}A.$$

In the matrix form is

$$T = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix} - \begin{pmatrix} \frac{1}{a_{11}} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{a_{nn}} \end{pmatrix} \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} = \left[t_{ij}\right] = \begin{cases} t_{ij} = 0, \ i = j, \\ t_{ij} = -\frac{a_{ij}}{a_{ii}}, \ i \neq j. \end{cases}.$$

So,

$$\|T\|_\infty = \max_i \sum_j |t_{ij}| = \max_i \sum_{i\neq j} |\frac{a_{ij}}{a_{ii}}|.$$

Since A is diagonal dominant, so

$$|a_{ii}| \ge \sum_{j\neq i} |a_{ij}|.$$

Therefore,

$$1 \ge \sum_{j\neq i} \frac{|a_{ij}|}{|a_{ii}|}.$$

Hence, $\|T\|_\infty < 1$ ◀

Problem A.13. *(Sample #35) Let $A = [a_{i,j}] \in \mathbb{C}^{n\times n}$ be invertible and $b \in \mathbb{C}^n$. Prove that the classical Gauss-Seidel iteration method for approximating the solution to $Ax = b$ is convergent, for any starting value x_0, if A is strictly diagonally dominant, i.e.*

$$|a_{i,i}| < \sum_{k\neq i} |a_{i,k}|, \ \forall\ i = 1, \cdots, n.$$

Solution. The Jacobi iteration scheme is as follows

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

This scheme can be rewritten as

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

We want to show If A is diagonal dominant , then $\|T_{GS}\| < 1$, then Jacobi Method convergences. From the definition of T, we know that T for Gauss-Seidel iteration Method is as follows

$$T_{GS} = -(L+D)^{-1}U.$$

Since A is diagonal dominant, so So,

$$|a_{ii}| - \sum_{j<i} |a_{ij}| \geq \sum_{j>i} |a_{ij}|,$$

which implies

$$\gamma = max_i \left\{ \frac{\sum_{j>i} |a_{ij}|}{|a_{ii}| - \sum_{j<i} |a_{ij}|} \right\} \leq 1.$$

Now, we will show $\|T_{GS}\| < \gamma$. Let $x \in \mathbb{C}^n$ and $y = Tx$, i.e.

$$y = T_{GS}x = -(L+D)^{-1}Ux.$$

Let i_0 be the index such that $\|y\|_\infty = |y_{i_0}|$, then we have

$$|((L+D)y)_{i_0}| = |(Ux)_{i_0}| = |\sum_{j>i_0} a_{i_0 j}x_j| \leq \sum_{j>i_0} |a_{i_0 j}||x_j| \leq \sum_{j>i_0} |a_{i_0 j}| \|x\|_\infty.$$

Moreover

$$|((L+D)y)_{i_0}| = |\sum_{j<i_0} a_{i_0 j}y_j + a_{i_0 i_0}y_j| \geq |a_{i_0 i_0}y_j| - |\sum_{j<i_0} a_{i_0 j}y_j| = |a_{i_0 i_0}| \|y\|_\infty - |\sum_{j<i_0} a_{i_0 j}y_j| \geq |a_{i_0 i_0}| \|y\|_\infty - \sum_{j<i_0} |a_{i_0 j}| \|y\|_\infty.$$

Therefore, from the above two equations, we have

$$|a_{i_0 i_0}| \|y\|_\infty - \sum_{j<i_0} |a_{i_0 j}| \|y\|_\infty \leq \sum_{j>i_0} |a_{i_0 j}| \|x\|_\infty,$$

which implies

$$\|y\|_\infty \leq \frac{\sum_{j>i_0} |a_{i_0 j}|}{|a_{i_0 i_0}| - \sum_{j<i_0} |a_{i_0 j}|} \|x\|_\infty.$$

So,

$$\|T_{GS}x\|_\infty \le \gamma \|x\|_\infty ,$$

which implies

$$\|T_{GS}\|_\infty \le \gamma < 1.$$

◀

Problem A.14. *(Sample #38) Let $A \in \mathbb{C}^{n\times n}$ be invertible and suppose $b \in \mathbb{C}^n_*$ satisfies $Ax = b$. let the perturbations $\delta x, \delta b \in \mathbb{C}^n$ satisfy $A\delta x = \delta b$, so that $A(x+\delta x) = b + \delta b$.*

1. *Prove the error (or perturbation) estimate*

$$\frac{1}{\kappa(A)}\frac{\|\delta b\|}{\|b\|} \le \frac{\|\delta x\|}{\|x\|} \le \kappa(A)\frac{\|\delta b\|}{\|b\|}.$$

2. *Show that for any invertible matrix A, the upper bound for $\frac{\|\delta b\|}{\|b\|}$ above can be attained for suitable choice of b and δb. (In other words, the upper bound is sharp.)*

Solution. 1. Since $Ax = b$ and $A\delta x = \delta b$, then $x = A^{-1}b$ and

$$\|\delta b\| = \|A\delta x\| \le \|A\|\|\delta x\|, \|x\| = \left\|A^{-1}b\right\| \le \left\|A^{-1}\right\|\|b\|.$$

Therefore

$$\frac{\|\delta b\|}{\|A\|} \le \|\delta x\|, \frac{1}{\left\|A^{-1}\right\|\|b\|} \le \frac{1}{\|x\|}.$$

Hence

$$\frac{1}{\kappa(A)}\frac{\|\delta b\|}{\|b\|} \le \frac{\|\delta x\|}{\|x\|}.$$

Similarly, since $Ax = b$ and $A\delta x = \delta b$, then $\delta x = A^{-1}\delta b$ and

$$\|b\| = \|Ax\| \le \|A\|\|x\|, \|\delta x\| = \left\|A^{-1}\delta b\right\| \le \left\|A^{-1}\right\|\|\delta b\|.$$

Therefore

$$\frac{1}{\|x\|} \le \frac{\|A\|}{\|b\|}.$$

Hence,

$$\frac{\|\delta x\|}{\|x\|} \le \kappa(A)\frac{\|\delta b\|}{\|b\|}.$$

So,

$$\frac{1}{\kappa(A)}\frac{\|\delta b\|}{\|b\|} \le \frac{\|\delta x\|}{\|x\|} \le \kappa(A)\frac{\|\delta b\|}{\|b\|}.$$

2. Since $Ax = b$ and $A\delta x = \delta b$, then $x = A^{-1}b$ and

$$\frac{1}{\|b\|} \leq \frac{\left\|A^{-1}\right\|}{\|x\|}, \|\delta b\| = \|A\delta x\| \leq \|A\|\|\delta x\|.$$

Hence,

$$\frac{\|\delta b\|}{\|b\|} \leq \kappa(A)\frac{\|\delta x\|}{\|x\|}$$

So the upper bound for $\frac{\|\delta b\|}{\|b\|}$ above can be attained for suitable choice of b and δb, since x and δx are dependent on b and δb, respectively.

◀

Problem A.15. *(Sample #39) Let $A \in \mathbb{R}^{n\times n}, b \in \mathbb{R}^n$. Suppose x and $\hat{x}$ solve $Ax = b$ and $(A+\delta A)\hat{x} = b + \delta b$, respectively. Assuming that $\left\|A^{-1}\right\|\|\delta A\| < 1$, show that*

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\kappa(A)}{1 - \kappa_2(A)\frac{\|\delta A\|_2}{\|A\|_2}}\left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|}\right)$$

where $\delta x = \hat{x} - x$.

Solution. Since $\left\|A^{-1}\right\|\|\delta A\| < 1$, then we have

$$\left\|A^{-1}\delta A\right\| \leq \left\|A^{-1}\right\|\|\delta A\| < 1.$$

Therefore,

$$\left\|\left(I - A^{-1}\delta A\right)^{-1}\right\| \leq \frac{1}{1 - \left\|A^{-1}\delta A\right\|}.$$
$$\left\|\left(I + A^{-1}\delta A\right)^{-1}\right\| \leq \frac{1}{1 - \left\|A^{-1}\delta A\right\|}.$$

$$\begin{aligned}
\delta x &= x + \delta x - x \\
&= (A+\delta A)^{-1}(b+\delta b) - A^{-1}b \\
&= (A+\delta A)^{-1}AA^{-1}(b+\delta b) - A^{-1}b \\
&= (A+\delta A)^{-1}\left(A^{-1}\right)^{-1}A^{-1}(b+\delta b) - A^{-1}b \\
&= (A^{-1}A + A^{-1}\delta A)^{-1}A^{-1}(b+\delta b) - A^{-1}b \\
&= (I + A^{-1}\delta A)^{-1}A^{-1}(b+\delta b) - A^{-1}b \\
&= (I + A^{-1}\delta A)^{-1}\left(A^{-1}(b+\delta b) - (I + A^{-1}\delta A)A^{-1}b\right) \\
&= (I + A^{-1}\delta A)^{-1}\left(A^{-1}\delta b - A^{-1}\delta AA^{-1}b\right)
\end{aligned}$$

Therefore,

$$\begin{aligned}
\|\delta x\| &\leq \frac{1}{1-\left\|A^{-1}\delta A\right\|}\left(\left\|A^{-1}\delta b\right\|+\left\|A^{-1}\delta A A^{-1}b\right\|\right)\\
&\leq \frac{1}{1-\left\|A^{-1}\delta A\right\|}\left(\left\|A^{-1}\right\|\|\delta b\|+\left\|A^{-1}\right\|\|\delta A\|\left\|A^{-1}b\right\|\right)\\
&= \frac{1}{1-\left\|A^{-1}\delta A\right\|}\left(\left\|A^{-1}\right\|\|\delta b\|+\left\|A^{-1}\right\|\|\delta A\|\|x\|\right)\\
&= \frac{\kappa(A)}{1-\left\|A^{-1}\delta A\right\|}\left(\frac{\|\delta b\|}{\|A\|}+\frac{\|\delta A\|\|x\|}{\|A\|}\right)
\end{aligned}$$

Dividing $\|x\|$ on both sides of the above equation yields

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\kappa(A)}{1-\left\|A^{-1}\delta A\right\|}\left(\frac{\|\delta b\|}{\|A\|\|x\|}+\frac{\|\delta A\|}{\|A\|}\right)$$

Since $\|b\| = \|Ax\| \leq \|A\|\|x\|$, then we have

$$\begin{aligned}
\frac{\|\delta x\|}{\|x\|} &\leq \frac{\kappa(A)}{1-\left\|A^{-1}\delta A\right\|}\left(\frac{\|\delta b\|}{\|b\|}+\frac{\|\delta A\|}{\|A\|}\right)\\
&\leq \frac{\kappa(A)}{1-\left\|A^{-1}\right\|_2\|\delta A\|_2}\left(\frac{\|\delta b\|}{\|b\|}+\frac{\|\delta A\|}{\|A\|}\right)\\
&\leq \frac{\kappa(A)}{1-\frac{\|A^{-1}\|_2\|A\|_2\|\delta A\|_2}{\|A\|_2}}\left(\frac{\|\delta b\|}{\|b\|}+\frac{\|\delta A\|}{\|A\|}\right)\\
&= \frac{\kappa(A)}{1-\kappa_2(A)\frac{\|\delta A\|_2}{\|A\|_2}}\left(\frac{\|\delta b\|}{\|b\|}+\frac{\|\delta A\|}{\|A\|}\right).
\end{aligned}$$

◀

Problem A.16. *(Sample #39) Let $A \in \mathbb{R}^{n\times n}, b \in \mathbb{R}^n$. Suppose x and $\hat{x}$ solve $Ax = b$ and $(A+\delta A)\hat{x} = b$, respectively. Assuming that $\left\|A^{-1}\right\|\|\delta A\| < 1$, show that*

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\kappa(A)}{1-\kappa_2(A)\frac{\|\delta A\|_2}{\|A\|_2}}\frac{\|\delta A\|}{\|A\|}.$$

where $\delta x = \hat{x} - x$.

Solution. Since $\left\|A^{-1}\right\|\|\delta A\| < 1$, then we have

$$\left\|A^{-1}\delta A\right\| \leq \left\|A^{-1}\right\|\|\delta A\| < 1.$$

Therefore,

$$\left\|\left(I - A^{-1}\delta A\right)^{-1}\right\| \leq \frac{1}{1-\left\|A^{-1}\delta A\right\|}.$$

$$\left\|\left(I + A^{-1}\delta A\right)^{-1}\right\| \leq \frac{1}{1-\left\|A^{-1}\delta A\right\|}.$$

$$\begin{aligned}
\delta x &= x+\delta x-x \\
&= (A+\delta A)^{-1}b - A^{-1}b \\
&= (A+\delta A)^{-1}AA^{-1}b - A^{-1}b \\
&= (A+\delta A)^{-1}\left(A^{-1}\right)^{-1}A^{-1}b - A^{-1}b \\
&= (A^{-1}A + A^{-1}\delta A)^{-1}A^{-1}b - A^{-1}b \\
&= (I + A^{-1}\delta A)^{-1}A^{-1}b - A^{-1}b \\
&= (I + A^{-1}\delta A)^{-1}\left(A^{-1}b - (I+A^{-1}\delta A)A^{-1}b\right) \\
&= (I + A^{-1}\delta A)^{-1}\left(-A^{-1}\delta A A^{-1}b\right)
\end{aligned}$$

Therefore,

$$\begin{aligned}
\|\delta x\| &\leq \frac{1}{1-\left\|A^{-1}\delta A\right\|}\left(\left\|A^{-1}\delta A A^{-1}b\right\|\right) \\
&\leq \frac{1}{1-\left\|A^{-1}\delta A\right\|}\left(\left\|A^{-1}\right\|\|\delta A\|\left\|A^{-1}b\right\|\right) \\
&= \frac{1}{1-\left\|A^{-1}\delta A\right\|}\left(\left\|A^{-1}\right\|\|\delta A\|\|x\|\right) \\
&= \frac{\kappa(A)}{1-\left\|A^{-1}\delta A\right\|}\left(\frac{\|\delta A\|\|x\|}{\|A\|}\right)
\end{aligned}$$

Dividing $\|x\|$ on both sides of the above equation yields

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\kappa(A)}{1-\left\|A^{-1}\delta A\right\|}\left(\frac{\|\delta A\|}{\|A\|}\right)$$

Since $\|b\| = \|Ax\| \leq \|A\|\|x\|$, then we have

$$\begin{aligned}
\frac{\|\delta x\|}{\|x\|} &\leq \frac{\kappa(A)}{1-\left\|A^{-1}\delta A\right\|}\left(\frac{\|\delta A\|}{\|A\|}\right) \\
&\leq \frac{\kappa(A)}{1-\left\|A^{-1}\right\|_2\|\delta A\|_2}\left(\frac{\|\delta A\|}{\|A\|}\right) \\
&\leq \frac{\kappa(A)}{1-\frac{\|A^{-1}\|_2\|A\|_2\|\delta A\|_2}{\|A\|_2}}\left(\frac{\|\delta A\|}{\|A\|}\right) \\
&= \frac{\kappa(A)}{1-\kappa_2(A)\frac{\|\delta A\|_2}{\|A\|_2}}\frac{\|\delta A\|}{\|A\|}.
\end{aligned}$$

◀

Problem A.17. *(Sample #40) Show that if λ is an eigenvalue of A^*A, where $A \in \mathbb{C}^{n\times n}$, then*

$$0 \leq \lambda \leq \|A^*\|\|A\|.$$

Problem A.18. *(Sample #41) Suppose $A \in \mathbb{C}^{n\times n}$ is invertible. Show that*

$$\kappa_2(A) = \sqrt{\frac{\lambda_n}{\lambda_1}},$$

where λ_n is the largest eigenvalue of $B := A^*A$, and λ_1 is the smallest eigenvalue of B.

Problem A.19. *(Sample #42) Suppose $A \in \mathbb{C}^{n\times n}$ and A is invertible. Prove that*

$$\kappa_2 \le \sqrt{\kappa_1(A)\kappa_\infty(A)}.$$

Solution.

Claim A.2.

$$\|A\|_2^2 \le \|A\|_1 \|A\|_\infty .$$

Proof.

$$\|A\|_2^2 = \rho(A)^2 = \lambda \le \|A\|_1 \|A^*\|_1 = \|A\|_1 \|A\|_\infty .$$

where λ is the eigenvalue of A^*A. ◀

Since $\kappa_2 = \|A\|_2 \left\|A^{-1}\right\|_2$, $\kappa_1 = \|A\|_1 \left\|A^{-1}\right\|_1$ and $\kappa_\infty = \|A\|_\infty \left\|A^{-1}\right\|_\infty$.

$$\|A\|_2 \left\|A^{-1}\right\|_2 \le \sqrt{\|A\|_1 \|A\|_\infty}\sqrt{\left\|A^{-1}\right\|_1 \left\|A^{-1}\right\|_\infty} \le \sqrt{\|A\|_1 \left\|A^{-1}\right\|_1 \|A\|_\infty \left\|A^{-1}\right\|_\infty} = \sqrt{\kappa_1(A)\kappa_\infty(A)}.$$

◀

Problem A.20. *(Sample #44) Suppose $A, B \in \mathbb{R}^{n\times n}$ and A is non-singular and B is singular. Prove that*

$$\frac{1}{\kappa(A)} \le \frac{\|A-B\|}{\|A\|},$$

where $\kappa(A) = \|A\| \cdot \left\|A^{-1}\right\|$, and $\|\cdot\|$ is an reduced matrix norm.

Solution. Since B is singular, then there exists a vector $x \neq 0$, s.t. $Bx = 0$. Since A is non-singular, then A^{-1} is also non-singular. Moreover, $A^{-1}Bx = 0$. Then, we have

$$x = x - A^{-1}Bx = (I - A^{-1}B)x.$$

So

$$\|x\| = \left\|(I - A^{-1}B)x\right\| \le \left\|A^{-1}A - A^{-1}B\right\| \|x\| \le \left\|A^{-1}\right\| \|A - B\| \|x\|.$$

Since $x \neq 0$, so

$$1 \le \left\|A^{-1}\right\| \|A - B\|.$$

$$\frac{1}{\left\|A^{-1}\right\| \|A\|} \le \frac{\|A-B\|}{\|A\|},$$

i.e.

$$\frac{1}{\kappa(A)} \le \frac{\|A-B\|}{\|A\|}.$$

◀

A.2 Numerical Solutions of Nonlinear Equations

Problem A.21. *(Sample #1) Let $\{x^n\}$ be a sequence generated by Newton's method. Suppose that the initial guess x^0 is well chosen so that this sequence converges to the exact solution x_*. Prove that if $f(x_*) = f'(x_*) = \cdots = f^{m-1}(x_*) = 0, f^m(x_*) \neq 0$, x^n converges linearly to x_* with*

$$\lim_{k\to\infty} \frac{e^{k+1}}{e^k} = \frac{m-1}{m}.$$

Solution. Newton's method scheme is read as follows

$$x^{k+1} = x^k - \frac{f(x^k)}{f'(x^k)}.$$

Let $e^k = x^k - x_*$, then

$$\begin{aligned} e^{k+1} &= x^{k+1} - x_* \\ &= x^k - \frac{f(x^k)}{f'(x^k)} - x_* \\ &= e^k - \frac{f(x^k)}{f'(x^k)}. \end{aligned}$$

Therefore,

$$\frac{e^{k+1}}{e^k} = 1 - \frac{f(x^k)}{e^k f'(x^k)}.$$

Since x^0 is well chosen so that this sequence converges to the exact solution x_*, therefore we have the Taylor expansion for $f(x^k), f'(x^k)$ at x_*, i.e.

$$\begin{aligned} f(x^k) &= f(x_*) + f'(x_*)e^k + \cdots + \frac{f^{(m-1)}(x_*)}{(m-1)!}\left(e^k\right)^{m-1} + \frac{f^{(m)}(\xi^k)}{m!}\left(e^k\right)^m \\ &= \frac{f^{(m)}(\xi^k)}{m!}\left(e^k\right)^m,\ \xi^k \in [x_*, x^k]. \\ f'(x^k) &= f'(x_*) + f''(x_*)e^k + \cdots + \frac{f^{(m-1)}(x_*)}{(m-2)!}\left(e^k\right)^{m-2} + \frac{f^{(m)}(\eta^k)}{(m-1)!}\left(e^k\right)^{m-1} \\ &= \frac{f^{(m)}(\eta^k)}{(m-1)!}\left(e^k\right)^{m-1},\ \eta^k \in [x_*, x^k]. \end{aligned}$$

Hence,

$$\lim_{k\to\infty} \frac{e^{k+1}}{e^k} = 1 - \frac{f(x^k)}{e^k f'(x^k)} = 1 - \frac{\frac{f^{(m)}(\xi^k)}{m!}\left(e^k\right)^m}{\frac{f^{(m)}(\eta^k)}{(m-1)!}\left(e^k\right)^{m-1} e^k} = \frac{m-1}{m} = 1 - \frac{1}{m}\frac{f^{(m)}(\xi^k)}{f^{(m)}(\eta^k)} = 1 - \frac{1}{m} = \frac{m-1}{m},$$

since when $k \to \infty$ then $\xi^k, \eta^k \to x_*$. ◀

Problem A.22. *(Sample #2) Let $\mathbf{f}:\Omega\subset\mathbb{R}^n\to\mathbb{R}^n$ be twice continuously differentiable. Suppose $\mathbf{x}^*\in\Omega$ is a solution of $\mathbf{f}(\mathbf{x})=0$, and the Jacobian matrix of $\mathbf{f}$, denoted $J_{\mathbf{f}}$, is invertible at $\mathbf{x}^*$.*

1. *Prove that if $\mathbf{x}^0\in\Omega$ is sufficiently close to $\mathbf{x}^*$, then the following iteration converges to $\mathbf{x}^*$:*

$$\mathbf{x}^{k+1}=\mathbf{x}^k-\left(J_{\mathbf{f}}(\mathbf{x}^0)\right)^{-1}\mathbf{f}(\mathbf{x}^k).$$

2. *Prove that the convergence is typically linear.*

Solution. ◂

Problem A.23. *(Sample #3) Let $a\in\mathbb{R}^n$ and $R>0$ be given. Suppose that $\mathbf{f}:\overline{B}(\mathbf{a},R)\to\mathbb{R}^n$, $f_i\in C^2(\overline{B}(\mathbf{a},R))$, for each $i=1,\cdots,n$. Suppose that there is a point $\xi\in\overline{B}(\mathbf{a},R)$, such that $\mathbf{f}(\xi)=0$, and that the Jacobian matrix $J_{\mathbf{f}}(\mathbf{x})$ is invertible, with estimate $\left\|[J_{\mathbf{f}}(\mathbf{x})]^{-1}\right\|_2\le\beta$, for any $\mathbf{x}\in\overline{B}(\mathbf{a},R)$. Prove that the sequence $\{\mathbf{x}_k\}$ defined by Newton's method,*

$$J_{\mathbf{f}}(\mathbf{x}_k)\,(\mathbf{x}_{k+1}-\mathbf{x}_k)=-\mathbf{f}(\mathbf{x}_k),$$

converges (at least) Linear to the root ξ as $k\to\infty$, provided $\mathbf{x}_0$ is sufficiently close to ξ.

Solution. ◂

Problem A.24. *(Sample #6) Assume that $f:\mathbb{R}\to\mathbb{R}, f\in C^2(\mathbb{R}), f'(x)>0$ for all $x\in\mathbb{R}$, and $f''(x)>0$, for all $x\in\mathbb{R}$.*

1. *Suppose that a root $\xi\in\mathbb{R}$ exists. Prove that it is unique. Exhibit a function satisfying the assumptions above that has no root.*
2. *Prove that for any starting guess $x_0\in\mathbb{R}$, Newton's method converges, and the convergence rate is quadratic.*

Solution. 1. Let x_1 and x_2 are the two different roots. So, $f(x_1)=f(x_2)=0$, then by Mean value theorem, we have that there exists $\eta\in[x_1,x_2]$, such $f(\eta)=0$ which contradicts $f'(x)>0$.

2. example $f(x)=e^x$.

3. Let x^* be the root of $f(x)$. From the Taylor expansion, we know

$$0=f(x^*)=f(x^k)+f'(x^k)(x^*-x^k)+\frac{1}{2}f''(\theta)(x^*-x^k)^2,$$

where θ is between x^* and x^k. Define $e^k=x^*-x^k$, then

$$0=f(x^*)=f(x^k)+f'(x^k)(e^k)+\frac{1}{2}f''(\theta)(e^k)^2.$$

so

$$\left[f'(x^k)\right]^{-1}f(x^k)=-(e^k)-\frac{1}{2}\left[f'(x^k)\right]^{-1}f''(\theta)(e^k)^2.$$

From the Newton's scheme, we have

$$\begin{cases}x^{k+1}=x^k-\left[f'(x^k)\right]^{-1}f(x^k)\\ x^*=x^*\end{cases}$$

So,

$$e^{k+1} = e^k + \left[f'(x^k)\right]^{-1} f(x^k) = -\frac{1}{2}\left[f'(x^k)\right]^{-1} f''(\theta)(e^k)^2,$$

i.e.

$$e^{k+1} = -\frac{f''(\theta)}{2\left[f'(x^k)\right]}(e^k)^2,$$

By assumption, there is a neighborhood of x, such that

$$\left|f''(z)\right| \le C_1, \quad \left|f'(z)\right| \le C_2,$$

Therefore,

$$\left|e^{k+1}\right| \le \frac{\left|f''(\theta)\right|}{\left|2\left[f'(x^k)\right]\right|}(e^k)^2 \le \frac{C_1}{2C_2}\left|e^k\right|^2.$$

This implies

$$\left|x^{k+1} - x^*\right| \le C\left|x^k - x^*\right|^2.$$

◀

Problem A.25. *(Sample #8) Consider the two-step Newton method*

$$y_k = x_k - \frac{f(x_k)}{f'(x_k)}, x_{k+1} = y_k - \frac{f(y_k)}{f'(x_k)}$$

for the solution of the equation $f(x) = 0$. Prove

1. *If the method converges, then*

$$\lim_{k\to\infty} \frac{x_{k+1} - x^*}{(y_k - x^*)(x_k - x^*)} = \frac{f''(x_k)}{f'(x_k)},$$

   where x^* is the solution.

2. *Prove the convergence is cubic, that is*

$$\lim_{k\to\infty} \frac{x_{k+1} - x^*}{(x_k - x^*)^3} = \frac{1}{2}\left(\frac{f''(x_k)}{f'(x_k)}\right).$$

3. *Would you say that this method is faster than Newton's method given that its convergence is cubic?*

Solution. 1. First, we will show that if $x_k \in [x-h, x+h]$, then $y_k \in [x-h, x+h]$. By Taylor expansion formula, we have

$$0 = f(x^*) = f(x_k) + f'(x_k)(x^* - x_k) + \frac{1}{2!}f''(\xi_k)(x^* - x_k)^2,$$

where ξ is between x and x_k. Therefore, we have

$$f(x_k) = -f'(x_k)(x^* - x_k) - \frac{1}{2!}f''(\xi_k)(x^* - x_k)^2.$$

Plugging the above equation to the first step of the Newton's method, we have

$$y_k = x_k + (x^* - x_k) + \frac{1}{2!}\frac{f''(\xi_k)}{f'(x_k)}(x^* - x_k)^2.$$

then

$$y_k - x^* = \frac{1}{2!}\frac{f''(\xi_k)}{f'(x_k)}(x^* - x_k)^2. \tag{214}$$

Therefore,

$$\left|y_k - x^*\right| = \left|\frac{1}{2!}\frac{f''(\xi_k)}{f'(x_k)}(x^* - x_k)^2\right| \leq \frac{1}{2}\left|\frac{f''(\xi_k)}{f'(x_k)}\right|\left|(x^* - x_k)\right|\left|(x^* - x_k)\right|.$$

Since we can choose the initial value very close to x^*, shah that

$$\left|\frac{f''(\xi)}{f'(x_k)}\right|\left|(x^* - x_k)\right| \leq 1$$

Then, we have that

$$\left|y_k - x^*\right| \leq \frac{1}{2}\left|(x^* - x_k)\right|.$$

Hence, we proved the result, that is to say, if $x_k \to x^*$, then $y_k, \xi_k \to x^*$.

2. Next, we will show if $x_k \in [x-h, x+h]$, then $x_{k+1} \in [x-h, x+h]$. From the second step of the Newton's Method, we have that

$$\begin{aligned} x_{k+1} - x^* &= y_k - x^* - \frac{f(y_k)}{f'(x_k)} \\ &= \frac{1}{f'(x_k)}\left((y_k - x^*)f'(x_k) - f(y_k)\right) \\ &= \frac{1}{f'(x_k)}\left[(y_k - x^*)(f'(x_k) - f'(x^*)) - f(y_k) + (y_k - x^*)f'(x^*)\right] \end{aligned}$$

By mean value theorem, we have there exists η_k between x^* and x_k, such that

$$f'(x_k) - f'(x^*) = f''\eta_k(x_k - x^*),$$

and by Taylor expansion formula, we have

$$\begin{aligned} f(y_k) &= f(x^*) + f'(x^*)(y_k - x^*) + \frac{(y_k - x^*)^2}{2}f''(\gamma_k) \\ &= f'(x^*)(y_k - x^*) + \frac{(y_k - x^*)^2}{2}f''(\gamma_k), \end{aligned}$$

where γ is between y_k and x^*. Plugging the above two equations to the second step of the Newton's method, we get

$$\begin{aligned} x_{k+1} - x^* &= \frac{1}{f'(x_k)}\left[f''\eta_k(x_k - x^*)(y_k - x^*) - f'(x^*)(y_k - x^*) - \frac{(y_k - x^*)^2}{2}f''(\gamma_k) + (y_k - x^*)f'(x^*)\right] \\ &= \frac{1}{f'(x_k)}\left[f''\eta_k(x_k - x^*)(y_k - x^*) - \frac{(y_k - x^*)^2}{2}f''(\gamma_k)\right]. \end{aligned} \tag{215}$$

Taking absolute values of the above equation, then we have

$$\begin{aligned}\left|x_{k+1}-x^*\right| &= \left|\frac{1}{f'(x_k)}\left[f''\eta_k(x_k-x^*)(y_k-x^*)-\frac{(y_k-x^*)^2}{2}f''(\gamma_k)\right]\right| \\ &\le A|x_k-x^*|\left|y_k-x^*\right|+\frac{A}{2}\left|y_k-x^*\right|\left|y_k-x^*\right| \\ &\le \frac{1}{2}|x_k-x^*|+\frac{1}{8}|x_k-x^*|=\frac{5}{8}|x_k-x^*|.\end{aligned}$$

Hence, we proved the result, that is to say, if $y_k \to x^*$, then $x_{k+1}, \eta_k, \gamma_k \to x^*$.

3. Finally, we will prove the convergence order is cubic. From (215), we can get that

$$\frac{x_{k+1}-x^*}{(x_k-x^*)(y_k-x^*)}=\frac{f''\eta_k}{f'(x_k)}-\frac{(y_k-x^*)f''(\gamma_k)}{2(x_k-x^*)f'(x_k)}.$$

By using (214), we have

$$\frac{x_{k+1}-x^*}{(x_k-x^*)(y_k-x^*)}=\frac{f''\eta_k}{f'(x_k)}-\frac{1}{4}\frac{f''(\xi_k)}{f'(x_k)}(x^*-x_k)\frac{f''(\gamma_k)}{f'(x_k)}.$$

Taking limits gives

$$\lim_{k\to\infty}\frac{x_{k+1}-x^*}{(x_k-x^*)(y_k-x^*)}=\frac{f''(x^*)}{f'(x^*)}.$$

By using (214) again, we have

$$\frac{1}{y_k-x^*}=\frac{2}{(x^*-x_k)^2}\frac{f'(x_k)}{f''(\xi_k)}.$$

Hence

$$\lim_{k\to\infty}\frac{x_{k+1}-x^*}{(x_k-x^*)^3}=\frac{1}{2}\left(\frac{f''(x^*)}{f'(x^*)}\right)^2.$$

◀

A.3 Numerical Solutions of ODEs

Problem A.26. *(Sample #1) Show that, if z is a non-zero complex number that iix on the boundary of the linear stability domain of the two-step BDF method*

$$y_{n+2}-\frac{4}{3}y_{n+1}+\frac{1}{3}y_n=\frac{2}{3}hf(x_{n+2},y_{n+2}),$$

then the real part of z must be positive. Thus deduce that this method is A-stable.

Solution. For our this problem

$$\rho(w):=\sum_{m=0}^{s}a_m w^m=-2+w+w^2 \quad \text{and} \quad \sigma(w):=\sum_{m=0}^{s}b_m w^m=1+w+w^2. \tag{216}$$

By making the substitution with $\xi = w - 1$ i.e. $w = \xi + 1$, then

$$\rho(w) := \sum_{m=0}^{s} a_m w^m = \xi^2 + 3\xi \quad \text{and} \quad \sigma(w) := \sum_{m=0}^{s} b_m w^m = \xi^2 + 3\xi + 3. \tag{217}$$

So,

$$\begin{aligned}
\rho(w) - \sigma(w) ln(w) &= \xi^2 + 3\xi - (3 + 3\xi + \xi^2)(\xi - \frac{\xi^2}{2} + \frac{\xi^3}{3} \cdots) \\
&= \begin{array}{ccccc} +3\xi & +\xi^2 & & & \\ -3\xi & -3\xi^2 & -\xi^3 & & \\ & +\frac{3}{2}\xi^2 & +\frac{3}{2}\xi^3 & +\frac{1}{2}\xi^4 & \\ & & -\xi^3 & -\xi^4 & -\frac{1}{3}\xi^5 \end{array} \\
&= -\frac{1}{2}\xi^2 + \mathcal{O}(\xi^3).
\end{aligned}$$

Therefore, by the theorem

$$\rho(w) - \sigma(w) ln(w) = -\frac{1}{2}\xi^2 + \mathcal{O}(\xi^3).$$

◀

A.4 Numerical Solutions of PDEs

Problem A.27. *(Sample #1) Let V be a Hilbert space with inner product $(\cdot,\cdot)_V$ and norm $\|v\|_V = \sqrt{(v,v)_V}$, $\forall v \in V$. Suppose $a : V \times V \to \mathbb{R}$ is a symmetric bilinear form that is*

- *continuous:*

$$\left|a(u,v)\right| \le \gamma \|u\|_V \|v\|_V, \exists \gamma > 0, \forall u, v \in V,$$

- *coercive:*

$$\alpha \|u\|_V^2 \le \left|a(u,v)\right|, \exists \alpha > 0, \forall u \in V,$$

Suppose $L : V \to \mathbb{R}$ is linear and bound, i.e.

$$\left|L(u)\right| \le \lambda \|u\|_V,$$

for some $\lambda > 0, \forall u \in V$. Let u satisfies

$$a(u,v) = L(v)$$

, for all $v \in V$.

1. *Galerkin approximation: Suppose that $S_h \subset V$ is finite dimensional. Prove that there exists a unique $u_h \in V$ that satisfies*

$$a(u,v) = L(v)$$

, for all $v \in S_h$.

2. *Prove that the Galerkin approximation is stable $\|u_h\| \le \frac{\lambda}{\alpha}$.*
3. *Prove Ceá's lemma:*

$$\|u - u_h\|_V \le \frac{\gamma}{\alpha} \inf_{w \in S_h} \|u - w\|_V.$$

Solution. 1. Lax-Milgram theorem.

2. let $u_h \in S_h$ be the Galerkin approximation, then we have

$$\alpha \|u_h\|_V^2 \leq \left|a(u_h, u_h)\right| = \left|L(u_h)\right| \leq \lambda \|u_h\|_V .$$

So, we have

$$\|u_h\|_V \leq \frac{\lambda}{\alpha}.$$

3. let $u_h, w \in S_h$ be the Galerkin approximation, then we have

$$\begin{aligned} a(u, v) &= L(v) \\ a(u_h, v) &= L(v) \end{aligned}$$

then we have the so called Galerkin Orthogonal $a(u - u_h, v) = 0$ for all $v \in V$. Then by coercivity

$$\begin{aligned} \alpha \|u - u_h\|_V^2 &\leq \left|a(u - u_h, u - u_h)\right| \\ &= \left|a(u - u_h, u - w) + a(u - u_h, w - u_h)\right| \\ &= \left|a(u - u_h, u - w)\right| \\ &\leq \gamma \|u - u_h\|_V \|u - w\|_V . \end{aligned}$$

therefore

$$\|u - u_h\|_V \leq \frac{\gamma}{\alpha} \|u - w\|_V .$$

Hence

$$\|u - u_h\|_V \leq \frac{\gamma}{\alpha} \inf_{w \in S_h} \|u - w\|_V .$$

◀

Problem A.28. *(Sample #3) Consider the Lax-Friedrichs scheme,*

$$u_j^{n+1} = \frac{1}{2}\left(u_{j-1}^n + u_{j+1}^n\right) - \frac{\mu}{2}\left(u_{j+1}^n - u_{j-1}^n\right), \quad \mu = \frac{as}{h},$$

for approximating solutions to the Cauchy problem for the advection equation

$$\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = 0$$

where $a > 0$. Here $h > 0$ is the space step size, and $s > 0$ us the time step size.

1. *Prove that, if $s = C_1 h$, where C_1 is fixed positive constant, then the local truncation error satisfies the estimate*

$$\left|T_l^n\right| \leq C_0(s + h)$$

, where $C_0 > 0$ us a constant independent of s and h.

2. *Use the von Neumann analysis to show that the Lax-Friedrichs scheme is stable provided the CFL condition $0 < \mu = \frac{as}{h} \leq 1$ holds. In other words, compute the amplification factor, $g(\xi)$, and show that $\left|g(\xi)\right| \leq 1$, for all values of ξ, provided $\mu \leq 1$.*

Solution. 1. Then the Lax-Friedrichs method for solving the above partial differential equation is given by:

$$\frac{u_j^{n+1} - \frac{1}{2}(u_{j+1}^n + u_{j-1}^n)}{\Delta t} + a\frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} = 0$$

Or, rewriting this to solve for the unknown u_j^{n+1},

$$u_j^{n+1} = \frac{1}{2}(u_{j+1}^n + u_{j-1}^n) - a\frac{\Delta t}{2\Delta x}(u_{j+1}^n - u_{j-1}^n)$$

i.e.

$$u_j^{n+1} = \frac{1}{2}\left(u_{j-1}^n + u_{j+1}^n\right) - \frac{\mu}{2}\left(u_{j+1}^n - u_{j-1}^n\right), \quad \mu = \frac{as}{h}.$$

Let $\bar{u}$ be the exact solution and $\bar{u}_j^n = \bar{u}(n\Delta t, j\Delta x)$. Then from Taylor Expansion, we have

$$\begin{aligned}
\bar{u}_j^{n+1} &= \bar{u}_j^n + \Delta t\frac{\partial}{\partial t}\bar{u}_j^n + \frac{1}{2}(\Delta t)^2\frac{\partial^2}{\partial t^2}\bar{u}(\xi_1, j\Delta x), \quad t_n \le \xi_1 \le t_{n+1},\\
\bar{u}_{j-1}^n &= \bar{u}_j^n - \Delta x\frac{\partial}{\partial x}\bar{u}_j^n + \frac{1}{2}(\Delta x)^2\frac{\partial^2}{\partial x^2}\bar{u}(n\Delta t, \xi_2), \quad x_{j-1} \le \xi_2 \le x_j,\\
\bar{u}_{j+1}^n &= \bar{u}_j^n + \Delta x\frac{\partial}{\partial x}\bar{u}_j^n + \frac{1}{2}(\Delta x)^2\frac{\partial^2}{\partial x^2}\bar{u}(n\Delta t, \xi_3), \quad x_j \le \xi_3 \le x_{j+1}.
\end{aligned}$$

Then the truncation error T^{n+1} of this scheme is

$$\begin{aligned}
\left|T^{n+1}\right| &= \left|\frac{u_i^{n+1} - \frac{1}{2}(u_{i+1}^n + u_{i-1}^n)}{\Delta t} + a\frac{u_{i+1}^n - u_{i-1}^n}{2\Delta x}\right|\\
&= C\frac{\mathcal{O}(s)^2 + \mathcal{O}(h)^2}{s}
\end{aligned}$$

If $s = C_1 h$, where C_1 is fixed positive constant, then the local truncation error

$$T^{n+1} = C_0(s + h).$$

2. By using the Fourier transform, i.e.

$$u_j^{n+1} = g(\xi)u_j^n, \quad u_j^n = e^{ij\Delta x\xi},$$

then we have

$$g(\xi)u_j^n = \frac{1}{2}\left(u_{j-1}^n + u_{j+1}^n\right) - \frac{\mu}{2}\left(u_{j+1}^n - u_{j-1}^n\right),$$

and

$$g(\xi)e^{ij\Delta x\xi} = \frac{1}{2}\left(e^{i(j-1)\Delta x\xi} + e^{i(j+1)\Delta x\xi}\right) - \frac{\mu}{2}\left(e^{i(j+1)\Delta x\xi} - e^{i(j-1)\Delta x\xi}\right).$$

Then we have

$$\begin{aligned}
g(\xi) &= \frac{1}{2}\left(e^{-i\Delta x\xi} + e^{i\Delta x\xi}\right) - \frac{\mu}{2}\left(e^{i\Delta x\xi} - e^{-i\Delta x\xi}\right)\\
&= \cos(\Delta x\xi) + i\mu\sin(\Delta x\xi).
\end{aligned}$$

From von Neumann analysis, we know that the Lax-Friedrichs scheme is stable if $\left|g(\xi)\right| \leq 1$, i.e.

$$(\cos(\Delta x\xi))^2 + (\mu \sin(\Delta x\xi))^2 \leq 1,$$

i.e.

$$\mu \leq 1.$$

◀

Problem A.29. *(Sample #4) Consider the linear reaction-diffusion problem*

$$\begin{cases} \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} - u & \text{for } 0 \leq x \leq 1,\ 0 \leq t \leq T \\ u(0,t) = u(1,t) = 0 & \text{for } 0 \leq t \leq T \\ u(x,0) = g(x) & \text{for } 0 \leq x \leq 1, \end{cases}$$

The Crank-Nicolson scheme for this problem is written as

$$u_j^{n+1} = u_j^n + \frac{\mu}{2}\left(u_{j-1}^{n+1} - 2u_j^{n+1} + u_{j+1}^{n+1} + u_{j-1}^n - 2u_j^n + u_{j+1}^n\right) - \frac{s}{2}\left(u_j^{n+1} + u_j^n\right)$$

where $\mu = \frac{s}{h^2}$. Prove that the method is stable in the sense that

$$\left\|u^{n+1}\right\|_\infty \leq \|u^n\|_\infty,$$

for all $n \geq 0$, if $0 < \mu + \frac{s}{2} \leq 1$.

Solution. This problem is similar to Sample #14. The scheme can be rewritten as

$$(1+\mu)u_j^{n+1} = \frac{\mu}{2}u_{j-1}^{n+1} - \frac{s}{2}u_j^{n+1} + \frac{\mu}{2}u_{j+1}^{n+1} + \frac{\mu}{2}u_{j-1}^n + (1-\mu-\frac{s}{2})u_j^n + \frac{\mu}{2}u_{j+1}^n.$$

Then we have

$$\left|1+\mu\right|\left|u_j^{n+1}\right| \leq \left|\frac{\mu}{2}\right|\left|u_{j-1}^{n+1}\right| + \left|\frac{s}{2}\right|\left|u_j^{n+1}\right| + \left|\frac{\mu}{2}\right|\left|u_{j+1}^{n+1}\right| + \left|\frac{\mu}{2}\right|\left|u_{j-1}^n\right| + \left|(1-\mu-\frac{s}{2})\right|\left|u_j^n\right| + \left|\frac{\mu}{2}\right|\left|u_{j+1}^n\right|.$$

Therefore

$$(1+\mu)\left\|u^{n+1}\right\|_\infty \leq \frac{\mu}{2}\left\|u^{n+1}\right\|_\infty + \frac{s}{2}\left\|u^{n+1}\right\|_\infty + \frac{\mu}{2}\left\|u^{n+1}\right\|_\infty + \frac{\mu}{2}\|u^n\|_\infty + \left|(1-\mu-\frac{s}{2})\right|\|u^n\|_\infty + \frac{\mu}{2}\|u^n\|_\infty.$$

if $0 < \mu + \frac{s}{2} \leq 1$, then

$$(1+\mu)\left\|u^{n+1}\right\|_\infty \leq \frac{\mu}{2}\left\|u^{n+1}\right\|_\infty + \frac{s}{2}\left\|u^{n+1}\right\|_\infty + \frac{\mu}{2}\left\|u^{n+1}\right\|_\infty + \frac{\mu}{2}\|u^n\|_\infty + (1-\mu-\frac{s}{2})\|u^n\|_\infty + \frac{\mu}{2}\|u^n\|_\infty.$$

i.e.

$$(1-\frac{s}{2})\left\|u^{n+1}\right\|_\infty \leq (1-\frac{s}{2})\|u^n\|_\infty$$

Hence

$$\left\|u^{n+1}\right\|_\infty \leq \|u^n\|_\infty.$$

◀

Problem A.30. *(Sample #8) 1D Discrete Poincaré inequality: Let $\Omega = (0,1)$ and Ω_h be a uniform grid of size h. If $Y \in \mathcal{U}_h$ is a mesh function on Ω_h such that $Y(0) = 0$, then there is a constant C, independent of Y and h, for which*

$$\|Y\|_{2,h} \leq C \left\|\bar{\delta} Y\right\|_{2,h}.$$

Solution. I consider the following uniform partition (Figure. A1) of the interval $(0,1)$ with N points.

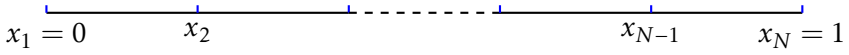

Figure A1: One dimension's uniform partition

Since the discrete 2-norm is defined as follows

$$\|v\|_{2,h}^2 = h^d \sum_{i=1}^{N} |v_i|^2,$$

where d is dimension. So, we have

$$\|v\|_{2,h}^2 = h \sum_{i=1}^{N} |v_i|^2, \quad \left\|\bar{\delta} v\right\|_{2,h}^2 = h \sum_{i=2}^{N} \left|\frac{v_{i-1} - v_i}{h}\right|^2.$$

Since $Y(0) = 0$, i.e. $Y_1 = 0$,

$$\sum_{i=2}^{N} (Y_{i-1} - Y_i) = Y_1 - Y_N = -Y_N.$$

Then,

$$\left|\sum_{i=2}^{N} Y_{i-1} - Y_i\right| = |Y_N|.$$

and

$$|Y_N| \leq \sum_{i=2}^{N} |Y_{i-1} - Y_i| = \sum_{i=2}^{N} h \left|\frac{Y_{i-1} - Y_i}{h}\right| \leq \left(\sum_{i=2}^{N} h^2\right)^{1/2} \left(\sum_{i=2}^{N} \left|\frac{Y_{i-1} - Y_i}{h}\right|^2\right)^{1/2}.$$

Therefore

$$\begin{aligned} |Y_K|^2 &\leq \left(\sum_{i=2}^{K} h^2\right)\left(\sum_{i=2}^{K} \left|\frac{Y_{i-1} - Y_i}{h}\right|^2\right) \\ &= h^2 (K-1) \sum_{i=2}^{K} \left|\frac{Y_{i-1} - Y_i}{h}\right|^2. \end{aligned}$$

1. When $K = 2$,

$$|Y_2|^2 \leq h^2 \left|\frac{Y_1 - Y_2}{h}\right|^2.$$

2. When $K = 3$,

$$|Y_3|^2 \leq 2h^2\left(\left|\frac{Y_1 - Y_2}{h}\right|^2 + \left|\frac{Y_2 - Y_3}{h}\right|^2\right).$$

3. When $K = N$,

$$|Y_N|^2 \leq (N-1)h^2\left(\left|\frac{Y_1 - Y_2}{h}\right|^2 + \left|\frac{Y_2 - Y_3}{h}\right|^2 + \cdots + \left|\frac{Y_{N-1} - Y_N}{h}\right|^2\right).$$

Sum over $|Y_i|^2$ from 2 to N, we get

$$\sum_{i=2}^{N} |Y_i|^2 \leq \frac{N(N-1)}{2} h^2 \sum_{i=2}^{N} \left|\frac{Y_{i-1} - Y_i}{h}\right|^2.$$

Since $Y_1 = 0$, so

$$\sum_{i=1}^{N} |Y_i|^2 \leq \frac{N(N-1)}{2} h^2 \sum_{i=2}^{N} \left|\frac{Y_{i-1} - Y_i}{h}\right|^2.$$

And then

$$\frac{1}{(N-1)^2} \sum_{i=1}^{N} |Y_i|^2 \leq \frac{N}{2(N-1)} h^2 \sum_{i=2}^{N} \left|\frac{Y_{i-1} - Y_i}{h}\right|^2 = \left(\frac{1}{2} + \frac{1}{2(N-1)}\right) h^2 \sum_{i=2}^{N} \left|\frac{Y_{i-1} - Y_i}{h}\right|^2.$$

Since $h = \frac{1}{N-1}$, so

$$h^2 \sum_{i=1}^{N} |Y_i|^2 \leq \left(\frac{1}{2} + \frac{1}{2(N-1)}\right) h^2 \sum_{i=2}^{N} \left|\frac{Y_{i-1} - Y_i}{h}\right|^2.$$

then

$$h \sum_{i=1}^{N} |Y_i|^2 \leq \left(\frac{1}{2} + \frac{1}{2(N-1)}\right) h \sum_{i=2}^{N} \left|\frac{Y_{i-1} - Y_i}{h}\right|^2.$$

i.e,

$$\|Y\|_{2,h}^2 \leq \left(\frac{1}{2} + \frac{1}{2(N-1)}\right) \left\|\bar{\delta} Y\right\|_{2,h}^2.$$

since $N \geq 2$, so

$$\|Y\|_{2,h}^2 \leq \left\|\bar{\delta} Y\right\|_{2,h}^2.$$

Hence,

$$\|Y\|_{2,h} \leq C \left\|\bar{\delta} Y\right\|_{2,h}.$$

◀

Problem A.31. *(Sample #12) Discrete maximum principle: Let $A = tridiag\{a_i, b_i, c_i\}_{i=1}^n \in \mathbb{R}^{n\times n}$ be a tridiagional matrix with the properties that*

$$b_i > 0, \quad a_i, c_i \leq 0, \quad a_i + b_i + c_i = 0.$$

Prove the following maximum principle: If $u \in \mathbb{R}^n$ is such that $(Au)_{i=2,\cdots,n-1} \leq 0$, then $u_i \leq \max\{u_1, u_n\}$.

Solution. Without loss generality, we assume $u_k, k = 2, \cdots, n-1$ is the maximum value.

1. For $(Au)_{i=2,\cdots,n-1} < 0$:
   I will use the method of contradiction to prove this case. Since $(Au)_{i=2,\cdots,n-1} < 0$, so
   $$a_k u_{k-1} + b_k u_k + c_k u_{k+1} < 0.$$
   Since $a_k + c_k = -b_k$ and $a_k < 0, c_k < 0$, so
   $$a_k u_{k-1} - (a_k + c_k)u_k + c_k u_{k+1} = a_k(u_{k-1} - u_k) + c_k(u_{k+1} - u_k) \geq 0.$$
   This is contradiction to $(Au)_{i=2,\cdots,n-1} < 0$. Therefore, If $u \in \mathbb{R}^n$ is such that $(Au)_{i=2,\cdots,n-1} < 0$, then $u_i \leq \max\{u_1, u_n\}$.
2. For $(Au)_{i=2,\cdots,n-1} = 0$:
   Since $(Au)_{i=2,\cdots,n-1} = 0$, so
   $$a_k u_{k-1} + b_k u_k + c_k u_{k+1} = 0.$$
   Since $a_k + c_k = -b_k$, so
   $$a_k u_{k-1} - (a_k + c_k)u_k + c_k u_{k+1} = a_k(u_{k-1} - u_k) + c_k(u_{k+1} - u_k) = 0.$$
   And $a_k < 0, c_k < 0, u_{k-1} - u_k \leq 0, u_{k+1} - u_k \leq 0$, so $u_{k-1} = u_k = u_{k+1}$, that is to say, u_{k-1} and u_{k+1} is also the maximum points. Bu using the same argument again, we get $u_{k-2} = u_{k-1} = u_k = u_{k+1} = u_{k+2}$. Repeating the process, we get
   $$u_1 = u_2 = \cdots = u_{n-1} = u_n.$$
   Therefore, If $u \in \mathbb{R}^n$ is such that $(Au)_{i=2,\cdots,n-1} = 0$, then $u_i \leq \max\{u_1, u_n\}$

◀

Problem A.32. *(Sample #14) Consider the Crank-Nicolson scheme*

$$u_j^{n+1} = u_j^n + \frac{\mu}{2}\left(u_{j-1}^{n+1} - 2u_j^{n+1} + u_{j+1}^{n+1} + u_{j-1}^n - 2u_j^n + u_{j+1}^n\right)$$

for approximating the solution to the heat equation $\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$ on the intervals $0 \leq x \leq 1$ and $0 \leq t \leq t^*$ with the boundary conditions $u(0,t) = u(1,t) = 0$.

1. *Show that the scheme may be written in the form $\mathbf{u}^{n+1} = A\mathbf{u}^n$, where $A \in \mathbb{R}_{sym}^{m\times m}$(the space of $m \times m$ symmetric matrices) and*
   $$\|Ax\|_2 \leq \|x\|_2,$$
   for any $\mathbf{x} \in \mathbb{R}^m$, regardless of the value of μ.
2. *Show that*
   $$\|Ax\|_\infty \leq \|x\|_\infty,$$
   for any $\mathbf{x} \in \mathbb{R}^m$, provided $\mu \leq 1$.(In other words, the scheme may only be conditionally stable in the max norm.)

Solution. 1. the scheme

$$u_j^{n+1} = u_j^n + \frac{\mu}{2}\left(u_{j-1}^{n+1} - 2u_j^{n+1} + u_{j+1}^{n+1} + u_{j-1}^n - 2u_j^n + u_{j+1}^n\right)$$

can be rewritten as

$$-\frac{\mu}{2}u_{j-1}^{n+1} + (1+\mu)u_j^{n+1} - \frac{\mu}{2}u_{j+1}^{n+1} = \frac{\mu}{2}u_{j-1}^n + (1-\mu)u_j^n + \frac{\mu}{2}u_{j+1}^n.$$

By using the boundary, we have

$$C\mathbf{u}^{n+1} = B\mathbf{u}^n$$

where

$$C = \begin{bmatrix} 1+\mu & -\frac{\mu}{2} & & & \\ -\frac{\mu}{2} & 1+\mu & -\frac{\mu}{2} & & \\ & \ddots & \ddots & \ddots & \\ & & -\frac{\mu}{2} & 1+\mu & -\frac{\mu}{2} \\ & & & -\frac{\mu}{2} & 1+\mu \end{bmatrix}, B = \begin{bmatrix} 1-\mu & \frac{\mu}{2} & & & \\ \frac{\mu}{2} & 1-\mu & \frac{\mu}{2} & & \\ & \ddots & \ddots & \ddots & \\ & & \frac{\mu}{2} & 1-\mu & \frac{\mu}{2} \\ & & & \frac{\mu}{2} & 1-\mu \end{bmatrix},$$

$$\mathbf{u}^{n+1} = \begin{bmatrix} u_1^{n+1} \\ u_2^{n+1} \\ \vdots \\ u_m^{n+1} \end{bmatrix} \text{ and } \mathbf{u}^n = \begin{bmatrix} u_1^n \\ u_2^n \\ \vdots \\ u_m^n \end{bmatrix}.$$

So, the scheme may be written in the form $\mathbf{u}^{n+1} = A\mathbf{u}^n$, where $A = C^{-1}B$. By using the Fourier transform, i.e.

$$u_j^{n+1} = g(\xi)u_j^n, \;\; u_j^n = e^{ij\Delta x\xi},$$

then we have

$$-\frac{\mu}{2}g(\xi)u_{j-1}^n + (1+\mu)g(\xi)u_j^n - \frac{\mu}{2}g(\xi)u_{j+1}^n = \frac{\mu}{2}u_{j-1}^n + (1-\mu)u_j^n + \frac{\mu}{2}u_{j+1}^n.$$

And then

$$-\frac{\mu}{2}g(\xi)e^{i(j-1)\Delta x\xi} + (1+\mu)g(\xi)e^{ij\Delta x\xi} - \frac{\mu}{2}g(\xi)e^{i(j+1)\Delta x\xi} = \frac{\mu}{2}e^{i(j-1)\Delta x\xi} + (1-\mu)e^{ij\Delta x\xi} + \frac{\mu}{2}e^{i(j+1)\Delta x\xi},$$

i.e.

$$g(\xi)\left(-\frac{\mu}{2}e^{-i\Delta x\xi} + (1+\mu) - \frac{\mu}{2}e^{i\Delta x\xi}\right)e^{j\Delta x\xi} = \left(\frac{\mu}{2}e^{-i\Delta x\xi} + (1-\mu) + \frac{\mu}{2}e^{i\Delta x\xi}\right)e^{j\Delta x\xi},$$

i.e.

$$g(\xi)\left(1+\mu-\mu\cos(\Delta x\xi)\right) = 1-\mu+\mu\cos(\Delta x\xi).$$

therefore,

$$g(\xi) = \frac{1-\mu+\mu\cos(\Delta x\xi)}{1+\mu-\mu\cos(\Delta x\xi)}.$$

hence

$$g(\xi) = \frac{1+\frac{1}{2}z}{1-\frac{1}{2}z}, z = 2\frac{\Delta t}{\Delta x^2}(cos(\Delta x\xi)-1).$$

Moreover, $\left|g(\xi)\right| < 1$, therefore, $\rho(A) < 1$.

$$\|Ax\|_2 \le \|A\|_2 \|x\|_2 = \rho(A)\|x\|_2 \le \|x\|_2.$$

2. the scheme

$$u_j^{n+1} = u_j^n + \frac{\mu}{2}\left(u_{j-1}^{n+1} - 2u_j^{n+1} + u_{j+1}^{n+1} + u_{j-1}^n - 2u_j^n + u_{j+1}^n\right)$$

can be rewritten as

$$(1+\mu)u_j^{n+1} = \frac{\mu}{2}u_{j-1}^{n+1} + \frac{\mu}{2}u_{j+1}^{n+1} + \frac{\mu}{2}u_{j-1}^n + (1-\mu)u_j^n + \frac{\mu}{2}u_{j+1}^n.$$

then

$$\left|1+\mu\right|\left|u_j^{n+1}\right| \le \left|\frac{\mu}{2}\right|\left|u_{j-1}^{n+1}\right| + \left|\frac{\mu}{2}\right|\left|u_{j+1}^{n+1}\right| + \left|\frac{\mu}{2}\right|\left|u_{j-1}^n\right| + \left|(1-\mu)\right|\left|u_j^n\right| + \left|\frac{\mu}{2}\right|\left|u_{j+1}^n\right|.$$

Therefore

$$(1+\mu)\left\|u_j^{n+1}\right\|_\infty \le \frac{\mu}{2}\left\|u_{j-1}^{n+1}\right\|_\infty + \frac{\mu}{2}\left\|u_{j+1}^{n+1}\right\|_\infty + \frac{\mu}{2}\left\|u_{j-1}^n\right\|_\infty + \left|(1-\mu)\right|\left\|u_j^n\right\|_\infty + \frac{\mu}{2}\left\|u_{j+1}^n\right\|_\infty.$$

i.e.

$$(1+\mu)\left\|\mathbf{u}^{n+1}\right\|_\infty \le \frac{\mu}{2}\left\|\mathbf{u}^{n+1}\right\|_\infty + \frac{\mu}{2}\left\|\mathbf{u}^{n+1}\right\|_\infty + \frac{\mu}{2}\|\mathbf{u}^n\|_\infty + \left|(1-\mu)\right|\|\mathbf{u}^n\|_\infty + \frac{\mu}{2}\|\mathbf{u}^n\|_\infty.$$

if $\mu \le 1$, then

$$\left\|\mathbf{u}^{n+1}\right\|_\infty \le \|\mathbf{u}^n\|_\infty,$$

i.e.

$$\|A\mathbf{u}^n\|_\infty \le \|\mathbf{u}^n\|_\infty.$$

◀

Problem A.33. *(Sample #15) Consider the Lax-Wendroff scheme*

$$u_j^{n+1} = u_j^n + \frac{a^2(\Delta t)^2}{2(\Delta x)^2}\left(u_{j-1}^n - 2u_j^n + u_{j+1}^n\right) - \frac{a\Delta t}{2\Delta x}\left(u_{j+1}^n - u_{j-1}^n\right),$$

for the approximating the solution of the Cauchy problem for the advection equation

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0, a > 0.$$

Use Von Neumann's Method to show that the Lax-Wendroff scheme is stable provided the CFL condition

$$\frac{a\Delta t}{\Delta x} \le 1.$$

is enforced.

Solution. By using the Fourier transform, i.e.

$$u_j^{n+1} = g(\xi)u_j^n, \quad u_j^n = e^{ij\Delta x\xi},$$

then we have

$$g(\xi)u_j^n = u_j^n + \frac{a^2(\Delta t)^2}{2(\Delta x)^2}\left(u_{j-1}^n - 2u_j^n + u_{j+1}^n\right) - \frac{a\Delta t}{2\Delta x}\left(u_{j+1}^n - u_{j-1}^n\right).$$

And then

$$g(\xi)e^{ij\Delta x\xi} = e^{ij\Delta x\xi} + \frac{a^2(\Delta t)^2}{2(\Delta x)^2}\left(e^{i(j-1)\Delta x\xi} - 2e^{ij\Delta x\xi} + e^{i(j+1)\Delta x\xi}\right) - \frac{a\Delta t}{2\Delta x}\left(e^{i(j+1)\Delta x\xi} - e^{i(j-1)\Delta x\xi}\right).$$

Therefore

$$\begin{aligned} g(\xi) &= 1 + \frac{a^2(\Delta t)^2}{2(\Delta x)^2}\left(e^{-i\Delta x\xi} - 2 + e^{i\Delta x\xi}\right) - \frac{a\Delta t}{2\Delta x}\left(e^{i\Delta x\xi} - e^{-i\Delta x\xi}\right) \\ &= 1 + \frac{a^2(\Delta t)^2}{2(\Delta x)^2}\left(2\cos(\Delta x\xi) - 2\right) - \frac{a\Delta t}{2\Delta x}\left(2i\sin(\Delta x\xi)\right) \\ &= 1 + \frac{a^2(\Delta t)^2}{(\Delta x)^2}\left(\cos(\Delta x\xi) - 1\right) - \frac{a\Delta t}{\Delta x}\left(i\sin(\Delta x\xi)\right). \end{aligned}$$

Let $\mu = \frac{a\Delta t}{\Delta x}$, then

$$g(\xi) = 1 + \mu^2\left(\cos(\Delta x\xi) - 1\right) - \mu\left(i\sin(\Delta x\xi)\right).$$

If $\left|g(\xi)\right| < 1$, then the scheme is stable, i,e

$$\left(1 + \mu^2\left(\cos(\Delta x\xi) - 1\right)\right)^2 + \left(\mu\sin(\Delta x\xi)\right)^2 < 1.$$

i.e.

$$1 + 2\mu^2\left(\cos(\Delta x\xi) - 1\right) + \mu^4\left(\cos(\Delta x\xi) - 1\right)^2 + \mu^2\sin(\Delta x\xi)^2 < 1.$$

i.e.

$$\mu^2\left(\sin(\Delta x\xi)^2 + 2\cos(\Delta x\xi) - 2\right) + \mu^4\left(\cos(\Delta x\xi) - 1\right)^2 < 0.$$

i.e.

$$\mu^2\left(1 - \cos(\Delta x\xi)^2 + 2\cos(\Delta x\xi) - 2\right) + \mu^4\left(\cos(\Delta x\xi) - 1\right)^2 < 0.$$

i.e

$$\mu^2(\cos(\Delta x\xi) - 1)^2 - (\cos(\Delta x\xi) - 1)^2 < 0,$$

$$(\mu^2 - 1)(\cos(\Delta x\xi) - 1)^2 < 0,$$

then we get $\mu < 1$. The above process is invertible, therefore, we prove the result. ◀

Problem A.34. *(Sample #16) Consider the Crank-Nicholson scheme applied to the diffusion equation*

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$$

where $t > 0, -\infty < x < \infty$.

1. *Show that the amplification factor in the Von Neumann analysis of the scheme us*

$$g(\xi) = \frac{1 + \frac{1}{2}z}{1 - \frac{1}{2}z}, z = 2\frac{\Delta t}{\Delta x^2}(cos(\Delta x\xi) - 1).$$

2. *Use the results of part 1 to show that the scheme is stable.*

Solution. 1. The Crank-Nicholson scheme for the diffusion equation is

$$\frac{u_j^{n+1}-u_j^n}{\Delta t}=\frac{1}{2}\left(\frac{u_{j-1}^{n+1}-2u_j^{n+1}+u_{j+1}^{n+1}}{\Delta x^2}+\frac{u_{j-1}^{n}-2u_j^{n}+u_{j+1}^{n}}{\Delta x^2}\right)$$

Let $\mu=\frac{\Delta t}{\Delta x^2}$, then the scheme can be rewrote as

$$u_j^{n+1}=u_j^n+\frac{\mu}{2}\left(u_{j-1}^{n+1}-2u_j^{n+1}+u_{j+1}^{n+1}+u_{j-1}^{n}-2u_j^{n}+u_{j+1}^{n}\right),$$

i.e.

$$-\frac{\mu}{2}u_{j-1}^{n+1}+(1+\mu)u_j^{n+1}-\frac{\mu}{2}u_{j+1}^{n+1}=\frac{\mu}{2}u_{j-1}^{n}+(1-\mu)u_j^{n}+\frac{\mu}{2}u_{j+1}^{n}.$$

By using the Fourier transform, i.e.

$$u_j^{n+1}=g(\xi)u_j^n,\ \ u_j^n=e^{ij\Delta x\xi},$$

then we have

$$-\frac{\mu}{2}g(\xi)u_{j-1}^{n}+(1+\mu)g(\xi)u_j^{n}-\frac{\mu}{2}g(\xi)u_{j+1}^{n}=\frac{\mu}{2}u_{j-1}^{n}+(1-\mu)u_j^{n}+\frac{\mu}{2}u_{j+1}^{n}.$$

And then

$$-\frac{\mu}{2}g(\xi)e^{i(j-1)\Delta x\xi}+(1+\mu)g(\xi)e^{ij\Delta x\xi}-\frac{\mu}{2}g(\xi)e^{i(j+1)\Delta x\xi}=\frac{\mu}{2}e^{i(j-1)\Delta x\xi}+(1-\mu)e^{ij\Delta x\xi}+\frac{\mu}{2}e^{i(j+1)\Delta x\xi},$$

i.e.

$$g(\xi)\left(-\frac{\mu}{2}e^{-i\Delta x\xi}+(1+\mu)-\frac{\mu}{2}e^{i\Delta x\xi}\right)e^{j\Delta x\xi}=\left(\frac{\mu}{2}e^{-i\Delta x\xi}+(1-\mu)+\frac{\mu}{2}e^{i\Delta x\xi}\right)e^{j\Delta x\xi},$$

i.e.

$$g(\xi)\left(1+\mu-\mu\cos(\Delta x\xi)\right)=1-\mu+\mu\cos(\Delta x\xi).$$

therefore,

$$g(\xi)=\frac{1-\mu+\mu\cos(\Delta x\xi)}{1+\mu-\mu\cos(\Delta x\xi)}.$$

hence

$$g(\xi)=\frac{1+\frac{1}{2}z}{1-\frac{1}{2}z},z=2\frac{\Delta t}{\Delta x^2}(cos(\Delta x\xi)-1).$$

2. since $z=2\frac{\Delta t}{\Delta x^2}(cos(\Delta x\xi)-1)$, then $z<0$, then we have

$$1+\frac{1}{2}z<1-\frac{1}{2}z,$$

therefore $g(\xi)<1$. Since $-1<1$, then

$$\frac{1}{2}z-1<\frac{1}{2}z+1.$$

Therefore,

$$g(\xi)=\frac{1+\frac{1}{2}z}{1-\frac{1}{2}z}>-1.$$

hence $\left|g(\xi)\right|<1$. So, the scheme is stable.

◀

Problem A.35. *(Sample #17) Consider the explicit scheme*

$$u_j^{n+1} = u_j^n + \mu\left(u_{j-1}^n - 2u_j^n + u_{j+1}^n\right) - \frac{b\mu\Delta x}{2}\left(u_{j+1}^n - u_{j-1}^n\right), 0 \le n \le N, 1 \le j \le L.$$

for the convention-diffusion problem

$$\begin{cases} \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} - b\frac{\partial u}{\partial x} & \text{for } 0 \le x \le 1,\ 0 \le t \le t^* \\ u(0,t) = u(1,t) = 0 & \text{for } 0 \le t \le t^* \\ u(x,0) = g(x) & \text{for } 0 \le x \le 1, \end{cases}$$

where $b > 0$, $\mu = \frac{\Delta t}{(\Delta x)^2}, \Delta x = \frac{1}{L+1}$, *and* $\Delta t = \frac{t^*}{N}$. *Prove that, under suitable restrictions on* μ *and* Δx, *the error grid function* e^n *satisfy the estimate*

$$\|e^n\|_\infty \le t^* C\left(\Delta t + \Delta x^2\right),$$

for all n *such that* $n\Delta t \le t^*$, *where* $C > 0$ *is a constant.*

Solution. Let $\bar{u}$ be the exact solution and $\bar{u}_j^n = \bar{u}(n\Delta t, j\Delta x)$. Then from Taylor Expansion, we have

$$\begin{aligned} \bar{u}_j^{n+1} &= \bar{u}_j^n + \Delta t\frac{\partial}{\partial t}\bar{u}_j^n + \frac{1}{2}(\Delta t)^2\frac{\partial^2}{\partial t^2}\bar{u}(\xi_1, j\Delta x), \quad t_n \le \xi_1 \le t_{n+1}, \\ \bar{u}_{j-1}^n &= \bar{u}_j^n - \Delta x\frac{\partial}{\partial x}\bar{u}_j^n - \frac{1}{6}(\Delta x)^3\frac{\partial^3}{\partial x^3}\bar{u}_j^n + \frac{1}{24}(\Delta x)^4\frac{\partial^4}{\partial x^4}\bar{u}(n\Delta t, \xi_2), \quad x_{j-1} \le \xi_2 \le x_j, \\ \bar{u}_{j+1}^n &= \bar{u}_j^n + \Delta x\frac{\partial}{\partial x}\bar{u}_j^n + \frac{1}{6}(\Delta x)^3\frac{\partial^3}{\partial x^3}\bar{u}_j^n + \frac{1}{24}(\Delta x)^4\frac{\partial^4}{\partial x^4}\bar{u}(n\Delta t, \xi_3), \quad x_j \le \xi_3 \le x_{j+1}. \end{aligned}$$

Then the truncation error T of this scheme is

$$\begin{aligned} T &= \frac{\bar{u}_j^{n+1} - \bar{u}_j^n}{\Delta t} - \frac{\bar{u}_{j-1}^n - 2\bar{u}_j^n + \bar{u}_{j+1}^n}{\Delta x^2} + b\frac{\bar{u}_{j+1}^n - \bar{u}_{j-1}^n}{\Delta x} \\ &= \mathcal{O}(\Delta t + (\Delta x)^2). \end{aligned}$$

Therefore

$$e_j^{n+1} = e_j^n + \mu\left(e_{j-1}^n - 2e_j^n + e_{j+1}^n\right) - \frac{b\mu\Delta x}{2}\left(e_{j+1}^n - e_{j-1}^n\right) + c\Delta t(\Delta t + (\Delta x)^2),$$

i.e.

$$e_j^{n+1} = \left(\mu + \frac{b\mu\Delta x}{2}\right)e_{j-1}^n + (1-2\mu)e_j^n + \left(\mu - \frac{b\mu\Delta x}{2}\right)e_{j+1}^n + c\Delta t(\Delta t + (\Delta x)^2).$$

Then

$$\left|e_j^{n+1}\right| \le \left|\mu + \frac{b\mu\Delta x}{2}\right|\left|e_{j-1}^n\right| + \left|(1-2\mu)\right|\left|e_j^n\right| + \left|\mu - \frac{b\mu\Delta x}{2}\right|\left|e_{j+1}^n\right| + c\Delta t(\Delta t + (\Delta x)^2).$$

Therefore

$$\left\|e_j^{n+1}\right\|_\infty \le \left|\mu + \frac{b\mu\Delta x}{2}\right|\left\|e_{j-1}^n\right\|_\infty + \left|(1-2\mu)\right|\left\|e_j^n\right\|_\infty + \left|\mu - \frac{b\mu\Delta x}{2}\right|\left\|e_{j+1}^n\right\|_\infty + c\Delta t(\Delta t + (\Delta x)^2).$$

$$\left\|e^{n+1}\right\|_\infty \le \left|\mu + \frac{b\mu\Delta x}{2}\right|\|e^n\|_\infty + \left|(1-2\mu)\right|\|e^n\|_\infty + \left|\mu - \frac{b\mu\Delta x}{2}\right|\|e^n\|_\infty + c\Delta t(\Delta t + (\Delta x)^2).$$

If $1-2\mu \geq 0$ and $\mu - \frac{b\mu\Delta x}{2} \geq 0$, i.e. $\mu \leq \frac{1}{2}$ and $1 - \frac{1}{2}b\Delta x > 0$, then

$$\begin{aligned} \left\|e^{n+1}\right\|_\infty &\leq \left(\mu + \frac{b\mu\Delta x}{2}\right)\|e^n\|_\infty + ((1-2\mu))\,\|e^n\|_\infty + \left(\mu - \frac{b\mu\Delta x}{2}\right)\|e^n\|_\infty + c\Delta t(\Delta t + (\Delta x)^2) \\ &= \|e^n\|_\infty + c\Delta t(\Delta t + (\Delta x)^2). \end{aligned}$$

Then

$$\begin{aligned} \|e^n\|_\infty &\leq \left\|e^{n-1}\right\|_\infty + c\Delta t(\Delta t + (\Delta x)^2) \\ &\leq \left\|e^{n-2}\right\|_\infty + c2\Delta t(\Delta t + (\Delta x)^2) \\ &\leq \qquad\qquad \vdots \\ &\leq \left\|e^0\right\|_\infty + cn\Delta t(\Delta t + (\Delta x)^2) \\ &= ct^*(\Delta t + (\Delta x)^2). \end{aligned}$$

◀

A.5 Supplemental Problems

B Numerical Mathematics Preliminary Examination

B.1 Numerical Mathematics Preliminary Examination Jan. 2011

Problem B.1. *(Prelim Jan. 2011#1) Consider a linear system $Ax = b$ with $A \in \mathbb{R}^{n\times n}$. Richardson's method is an iterative method*

$$Mx^{k+1} = Nx^k + b$$

with $M = \frac{1}{w}, N = M - A = \frac{1}{w}I - A$, where w is a damping factor chosen to make M approximate A as well as possible. Suppose A is positive definite and $w > 0$. Let λ_1 and λ_n denote the smallest and largest eigenvalue of A.

1. *Prove that Richardson's method converges if and only if $w < \frac{2}{\lambda_n}$.*
2. *Prove that the optimal value of w is $w_0 = \frac{2}{\lambda_1+\lambda_n}$.*

Solution. 1. Since $M = \frac{1}{w}, N = M - A = \frac{1}{w}I - A$, then we have

$$x^{k+1} = (I - wA)x^k + bw.$$

So $T_R = I - wA$, From the sufficient and & necessary condition for convergence, we should have $\rho(T_R) < 1$. Since λ_i are the eigenvalue of A, then we have $1 - \lambda_i w$ are the eigenvalues of T_R. Hence Richardson's method converges if and only if $|1 - \lambda_i w| < 1$, i.e

$$-1 < 1 - \lambda_n w < \cdots < 1 - \lambda_1 w < 1,$$

i.e. $w < \frac{2}{\lambda_n}$.

2. the minimal attaches at $|1 - \lambda_n w| = |1 - \lambda_1 w|$ (Figure. B2), i.e

$$\lambda_n w - 1 = 1 - \lambda_1 w,$$

i,e

$$w_0 = \frac{2}{\lambda_1 + \lambda_n}.$$

◀

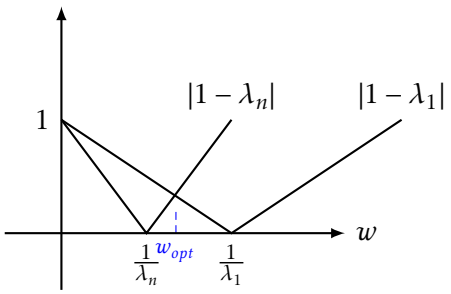

Figure B2: The curve of $\rho(T_R)$ as a function of w

Problem B.2. *(Prelim Jan. 2011#2) Let $A \in \mathbb{C}^{m\times n}$ and $b \in \mathbb{C}^m$. Prove that the vector $x \in \mathbb{C}^n$ is a least squares solution of $Ax = b$ if and only if $r \perp \text{range}(A)$, where $r = b - Ax$.*

Solution. We already know, $x \in \mathbb{C}^n$ is a least squares solution of $Ax = b$ if and only if

$$A^*Ax = A^*b.$$

and

$$\begin{aligned}(r, Ax) = (Ax) * r &= x^*A^*(b - Ax) \\ &= x^*(A^*b - A^*Ax)) \\ &= 0.\end{aligned}$$

Therefore, $r \perp \text{range}(A)$. The above way is invertible, hence we prove the result. ◀

Problem B.3. *(Prelim Jan. 2011#3) Suppose $A, B \in \mathbb{R}^{n\times n}$ and A is non-singular and B is singular. Prove that*

$$\frac{1}{\kappa(A)} \leq \frac{\|A - B\|}{\|A\|},$$

where $\kappa(A) = \|A\| \cdot \left\|A^{-1}\right\|$, and $\|\cdot\|$ is an reduced matrix norm.

Solution. Since B is singular, then there exists a vector $x \neq 0$, s.t. $Bx = 0$. Since A is non-singular, then A^{-1} is also non-singular. Moreover, $A^{-1}Bx = 0$. Then, we have

$$x = x - A^{-1}Bx = (I - A^{-1}B)x.$$

So

$$\|x\| = \left\|(I - A^{-1}B)x\right\| \leq \left\|A^{-1}A - A^{-1}B\right\| \|x\| \leq \left\|A^{-1}\right\| \|A - B\| \|x\|.$$

Since $x \neq 0$, so

$$1 \leq \left\|A^{-1}\right\| \|A - B\|.$$

$$\frac{1}{\left\|A^{-1}\right\| \|A\|} \leq \frac{\|A - B\|}{\|A\|},$$

i.e.

$$\frac{1}{\kappa(A)} \leq \frac{\|A - B\|}{\|A\|}.$$

◀

Problem B.4. *(Prelim Jan. 2011#4) Let $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}^n$ be twice continuously differentiable. Suppose $x^* \in \Omega$ is a solution of $f(x) = 0$, and the Jacobian matrix of f, denoted J_f, is invertible at x^*.*

1. *Prove that if $x^0 \in \Omega$ is sufficiently close to x^*, then the following iteration converges to x^*:*

$$x^{k+1} = x^k - J_f(x^0)^{-1} f(x^k).$$

2. *Prove that the convergence is typically only linear.*

Solution. Let $\mathbf{x}^*$ be the root of $\mathbf{f}(x)$ i.e. $\mathbf{f}(\mathbf{x}^*)$=0. From the Newton's scheme, we have

$$\begin{cases} \mathbf{x}^{k+1} = \mathbf{x}^k - \left[J(\mathbf{x}^0)\right]^{-1}\mathbf{f}(\mathbf{x}^k) \\ \mathbf{x}^* = \mathbf{x}^* \end{cases}$$

Therefore, we have

$$\begin{aligned} \mathbf{x}^* - \mathbf{x}^{k+1} &= \mathbf{x}^* - \mathbf{x}^k + \left[J(\mathbf{x}^0)\right]^{-1}(\mathbf{f}(\mathbf{x}^k) - \mathbf{f}(\mathbf{x}^*)) \\ &= \mathbf{x}^* - \mathbf{x}^k - \left[\mathbf{J}(\mathbf{x}^0)\right]^{-1}\mathbf{J}(\xi)(\mathbf{x}^* - \mathbf{x}^k). \end{aligned}$$

therefore

$$\left|\mathbf{x}^* - \mathbf{x}^{k+1}\right| \le \left|1 - \frac{\mathbf{J}(\xi)}{\mathbf{J}(\mathbf{x}^0)}\right|\left|\mathbf{x}^* - \mathbf{x}^k\right|$$

From theorem

Theorem B.1. *Suppose $J : \mathbb{R}^m \to \mathbb{R}^{n\times n}$ is a continuous matrix-valued function. If $J(x^*)$ is nonsingular, then there exists $\delta > 0$ such that, for all $x \in \mathrm{R}^m$ with $\|x - x^*\| < \delta$, $J(x)$ is nonsingular and*

$$\left\|J(x)^{-1}\right\| < 2\left\|J(x^*)^{-1}\right\|.$$

we get

$$\left|\mathbf{x}^* - \mathbf{x}^{k+1}\right| \le \frac{1}{2}\left|\mathbf{x}^* - \mathbf{x}^k\right|.$$

Which also shows the convergence is typically only linear.

◀

Problem B.5. *(Prelim Jan. 2011#5) Consider*

$$y'(t) = f(t, y(t)), \quad t \ge t_0, y(t_0) = y_0,$$

where $f : [t_0, t^*] \times \mathbb{R} \to \mathbb{R}$ is continuous in its first variable and Lipschitz continuous in its second variable. Prove that Euler's method converges.

Solution. The Euler's scheme is as follows:

$$y_{n+1} = y_n + hf(t_n, y_n), \qquad n = 0, 1, 2, \cdots. \tag{218}$$

By the Taylor expansion,

$$y(t_{n+1}) = y(t_n) + hy'(t_n) + \mathcal{O}(h^2).$$

So,

$$\begin{aligned} y(t_{n+1}) - y(t_n) - hf(t_n, y(t_n)) &= y(t_n) + hy'(t_n) + \mathcal{O}(h^2) - y(t_n) - hf(t_n, y(t_n)) \\ &= y(t_n) + hy'(t_n) + \mathcal{O}(h^2) - y(t_n) - hy'(t_n) \\ &= \mathcal{O}(h^2). \end{aligned} \tag{219}$$

Therefore, Forward Euler Method is order of 1 .

From (219), we get

$$y(t_{n+1}) = y(t_n) + hf(t_n, y(t_n)) + \mathcal{O}(h^2), \tag{220}$$

Subtracting (220) from (218), we get

$$e_{n+1} = e_n + h[f(t_n, y_n) - f(t_n, y(t_n))] + ch^2.$$

Since f is lipschitz continuous w.r.t. the second variable, then

$$|f(t_n, y_n) - f(t_n, y(t_n))| \le \lambda |y_n - y(t_n)|, \quad \lambda > 0.$$

Therefore,

$$\begin{aligned} \|e_{n+1}\| &\le \|e_n\| + h\lambda \|e_n\| + ch^2 \\ &= (1 + h\lambda)\|e_n\| + ch^2. \end{aligned}$$

Claim:[2]

$$\|e_n\| \le \frac{c}{\lambda} h[(1 + h\lambda)^n - 1], n = 0, 1, \cdots$$

Proof for Claim (221): The proof is by induction on n.

1. when $n = 0$, $e_n = 0$, hence $\|e_n\| \le \frac{c}{\lambda} h[(1 + h\lambda)^n - 1]$,
2. Induction assumption:

$$\|e_n\| \le \frac{c}{\lambda} h[(1 + h\lambda)^n - 1]$$

3. Induction steps:

$$\begin{aligned} \|e_{n+1}\| &\le (1 + h\lambda)\|e_n\| + ch^2 \\ &\le (1 + h\lambda)\frac{c}{\lambda} h[(1 + h\lambda)^n - 1] + ch^2 \\ &= \frac{c}{\lambda} h[(1 + h\lambda)^{n+1} - 1]. \end{aligned}$$

So, from the claim (221), we get $\|e_n\| \to 0$, when $h \to 0$. Therefore Forward Euler Method is convergent . ◀

Problem B.6. *(Prelim Jan. 2011#6) Consider the scheme*

$$y_{n+2} + y_{n+1} - 2y_n = h(f(t_{n+2}, y_{n+2}) + f(t_{n+1}, y_{n+1}) + f(t_n, y_n))$$

for approximating the solution to

$$y'(t) = f(t, y(t)), \quad t \ge t_0, y(t_0) = y_0,$$

what's the order of the scheme? Is it a convergent scheme? Is it A-stable? Justify your answers.

Solution. For our this problem

$$\rho(w) := \sum_{m=0}^{s} a_m w^m = -2 + w + w^2 \quad \text{and} \quad \sigma(w) := \sum_{m=0}^{s} b_m w^m = 1 + w + w^2. \tag{221}$$

By making the substitution with $\xi = w - 1$ i.e. $w = \xi + 1$, then

$$\rho(w) := \sum_{m=0}^{s} a_m w^m = \xi^2 + 3\xi \quad \text{and} \quad \sigma(w) := \sum_{m=0}^{s} b_m w^m = \xi^2 + 3\xi + 3. \tag{222}$$

So,

$$\begin{aligned}\rho(w)-\sigma(w)ln(w) &= \xi^2+3\xi-(3+3\xi+\xi^2)(\xi-\frac{\xi^2}{2}+\frac{\xi^3}{3}\cdots)\\ &= \begin{array}{cccc} +3\xi & +\xi^2 & & \\ -3\xi & -3\xi^2 & -\xi^3 & \\ & +\frac{3}{2}\xi^2 & +\frac{3}{2}\xi^3 & +\frac{1}{2}\xi^4 \\ & & -\xi^3 & -\xi^4 \quad -\frac{1}{3}\xi^5 \end{array}\\ &= -\frac{1}{2}\xi^2+\mathcal{O}(\xi^3).\end{aligned}$$

Therefore, by the theorem

$$\rho(w)-\sigma(w)ln(w)=-\frac{1}{2}\xi^2+\mathcal{O}(\xi^3).$$

Hence, this scheme is order of 1. Since,

$$\rho(w):=\sum_{m=0}^{s}a_m w^m=-2+w+w^2=(w+2)(w-1). \tag{223}$$

And $w=-1$ or $w=-2$ which does not satisfy the root condition. Therefore, this scheme is not stable. Hence, it is also not A-stable. ◀

Problem B.7. *(Prelim Jan. 2011#7) Consider the Crank-Nicholson scheme applied to the diffusion equation*

$$\frac{\partial u}{\partial t}=\frac{\partial^2 u}{\partial x^2}$$

where $t>0,-\infty<x<\infty$.

1. *Show that the amplification factor in the Von Neumann analysis of the scheme us*

$$g(\xi)=\frac{1+\frac{1}{2}z}{1-\frac{1}{2}z}, z=2\frac{\Delta t}{\Delta x^2}(cos(\Delta x\xi)-1).$$

2. *Use the results of part 1 to show that the scheme is stable.*

Solution. 1. The Crank-Nicholson scheme for the diffusion equation is

$$\frac{u_j^{n+1}-u_j^n}{\Delta t}=\frac{1}{2}\left(\frac{u_{j-1}^{n+1}-2u_j^{n+1}+u_{j+1}^{n+1}}{\Delta x^2}+\frac{u_{j-1}^{n}-2u_j^{n}+u_{j+1}^{n}}{\Delta x^2}\right)$$

Let $\mu=\frac{\Delta t}{\Delta x^2}$, then the scheme can be rewrote as

$$u_j^{n+1}=u_j^n+\frac{\mu}{2}\left(u_{j-1}^{n+1}-2u_j^{n+1}+u_{j+1}^{n+1}+u_{j-1}^{n}-2u_j^{n}+u_{j+1}^{n}\right),$$

i.e.

$$-\frac{\mu}{2}u_{j-1}^{n+1}+(1+\mu)u_j^{n+1}-\frac{\mu}{2}u_{j+1}^{n+1}=\frac{\mu}{2}u_{j-1}^{n}+(1-\mu)u_j^{n}+\frac{\mu}{2}u_{j+1}^{n}.$$

By using the Fourier transform, i.e.

$$u_j^{n+1}=g(\xi)u_j^n,\ \ u_j^n=e^{ij\Delta x\xi},$$

then we have

$$-\frac{\mu}{2}g(\xi)u_{j-1}^n+(1+\mu)g(\xi)u_j^n-\frac{\mu}{2}g(\xi)u_{j+1}^n=\frac{\mu}{2}u_{j-1}^n+(1-\mu)u_j^n+\frac{\mu}{2}u_{j+1}^n.$$

And then

$$-\frac{\mu}{2}g(\xi)e^{i(j-1)\Delta x\xi}+(1+\mu)g(\xi)e^{ij\Delta x\xi}-\frac{\mu}{2}g(\xi)e^{i(j+1)\Delta x\xi}=\frac{\mu}{2}e^{i(j-1)\Delta x\xi}+(1-\mu)e^{ij\Delta x\xi}+\frac{\mu}{2}e^{i(j+1)\Delta x\xi},$$

i.e.

$$g(\xi)\left(-\frac{\mu}{2}e^{-i\Delta x\xi}+(1+\mu)-\frac{\mu}{2}e^{i\Delta x\xi}\right)e^{j\Delta x\xi}=\left(\frac{\mu}{2}e^{-i\Delta x\xi}+(1-\mu)+\frac{\mu}{2}e^{i\Delta x\xi}\right)e^{j\Delta x\xi},$$

i.e.

$$g(\xi)\left(1+\mu-\mu\cos(\Delta x\xi)\right)=1-\mu+\mu\cos(\Delta x\xi).$$

therefore,

$$g(\xi)=\frac{1-\mu+\mu\cos(\Delta x\xi)}{1+\mu-\mu\cos(\Delta x\xi)}.$$

hence

$$g(\xi)=\frac{1+\frac{1}{2}z}{1-\frac{1}{2}z},z=2\frac{\Delta t}{\Delta x^2}(cos(\Delta x\xi)-1).$$

2. since $z=2\frac{\Delta t}{\Delta x^2}(cos(\Delta x\xi)-1)$, then $z<0$, then we have

$$1+\frac{1}{2}z<1-\frac{1}{2}z,$$

therefore $g(\xi)<1$. Since $-1<1$, then

$$\frac{1}{2}z-1<\frac{1}{2}z+1.$$

Therefore,

$$g(\xi)=\frac{1+\frac{1}{2}z}{1-\frac{1}{2}z}>-1.$$

hence $\left|g(\xi)\right|<1$. So, the scheme is stable.

◀

Problem B.8. *(Prelim Jan. 2011#8) Consider the explicit scheme*

$$u_j^{n+1} = u_j^n + \mu\left(u_{j-1}^n - 2u_j^n + u_{j+1}^n\right) - \frac{b\mu\Delta x}{2}\left(u_{j+1}^n - u_{j-1}^n\right), 0 \le n \le N, 1 \le j \le L.$$

for the convention-diffusion problem

$$\begin{cases} \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} - b\frac{\partial u}{\partial x} & \text{for } 0 \le x \le 1,\ 0 \le t \le t^* \\ u(0,t) = u(1,t) = 0 & \text{for } 0 \le t \le t^* \\ u(x,0) = g(x) & \text{for } 0 \le x \le 1, \end{cases}$$

where $b > 0$, $\mu = \frac{\Delta t}{(\Delta x)^2}, \Delta x = \frac{1}{L+1}$, and $\Delta t = \frac{t^*}{N}$. Prove that, under suitable restrictions on μ and Δx, the error grid function e^n satisfy the estimate

$$\|e^n\|_\infty \le t^* C\left(\Delta t + \Delta x^2\right),$$

for all n such that $n\Delta t \le t^*$, where $C > 0$ is a constant.

Solution. Let $\bar{u}$ be the exact solution and $\bar{u}_j^n = \bar{u}(n\Delta t, j\Delta x)$. Then from Taylor Expansion, we have

$$\begin{aligned}
\bar{u}_j^{n+1} &= \bar{u}_j^n + \Delta t\frac{\partial}{\partial t}\bar{u}_j^n + \frac{1}{2}(\Delta t)^2\frac{\partial^2}{\partial t^2}\bar{u}(\xi_1, j\Delta x), \quad t_n \le \xi_1 \le t_{n+1}, \\
\bar{u}_{j-1}^n &= \bar{u}_j^n - \Delta x\frac{\partial}{\partial x}\bar{u}_j^n + \frac{1}{2}(\Delta x)^2\frac{\partial^2}{\partial x^2}\bar{u}_j^n - \frac{1}{6}(\Delta x)^3\frac{\partial^3}{\partial x^3}\bar{u}_j^n + \frac{1}{24}(\Delta x)^4\frac{\partial^4}{\partial x^4}\bar{u}(n\Delta t, \xi_2), \quad x_{j-1} \le \xi_2 \le x_j, \\
\bar{u}_{j+1}^n &= \bar{u}_j^n + \Delta x\frac{\partial}{\partial x}\bar{u}_j^n + \frac{1}{2}(\Delta x)^2\frac{\partial^2}{\partial x^2}\bar{u}_j^n + \frac{1}{6}(\Delta x)^3\frac{\partial^3}{\partial x^3}\bar{u}_j^n + \frac{1}{24}(\Delta x)^4\frac{\partial^4}{\partial x^4}\bar{u}(n\Delta t, \xi_3), \quad x_j \le \xi_3 \le x_{j+1}.
\end{aligned}$$

Then the truncation error T of this scheme is

$$\begin{aligned}
T &= \frac{\bar{u}_j^{n+1} - \bar{u}_j^n}{\Delta t} - \frac{\bar{u}_{j-1}^n - 2\bar{u}_j^n + \bar{u}_{j+1}^n}{\Delta x^2} + b\frac{\bar{u}_{j+1}^n - \bar{u}_{j-1}^n}{\Delta x} \\
&= \mathcal{O}(\Delta t + (\Delta x)^2).
\end{aligned}$$

Therefore

$$e_j^{n+1} = e_j^n + \mu\left(e_{j-1}^n - 2e_j^n + e_{j+1}^n\right) - \frac{b\mu\Delta x}{2}\left(e_{j+1}^n - e_{j-1}^n\right) + c\Delta t(\Delta t + (\Delta x)^2),$$

i.e.

$$e_j^{n+1} = \left(\mu + \frac{b\mu\Delta x}{2}\right)e_{j-1}^n + (1-2\mu)e_j^n + \left(\mu - \frac{b\mu\Delta x}{2}\right)e_{j+1}^n + c\Delta t(\Delta t + (\Delta x)^2).$$

Then

$$\left|e_j^{n+1}\right| \le \left|\mu + \frac{b\mu\Delta x}{2}\right|\left|e_{j-1}^n\right| + \left|(1-2\mu)\right|\left|e_j^n\right| + \left|\mu - \frac{b\mu\Delta x}{2}\right|\left|e_{j+1}^n\right| + c\Delta t(\Delta t + (\Delta x)^2).$$

Therefore

$$\left\|e_j^{n+1}\right\|_\infty \le \left|\mu + \frac{b\mu\Delta x}{2}\right|\left\|e_{j-1}^n\right\|_\infty + \left|(1-2\mu)\right|\left\|e_j^n\right\|_\infty + \left|\mu - \frac{b\mu\Delta x}{2}\right|\left\|e_{j+1}^n\right\|_\infty + c\Delta t(\Delta t + (\Delta x)^2).$$

$$\left\|e^{n+1}\right\|_\infty \le \left|\mu + \frac{b\mu\Delta x}{2}\right|\|e^n\|_\infty + \left|(1-2\mu)\right|\|e^n\|_\infty + \left|\mu - \frac{b\mu\Delta x}{2}\right|\|e^n\|_\infty + c\Delta t(\Delta t + (\Delta x)^2).$$

If $1-2\mu \geq 0$ and $\mu - \frac{b\mu\Delta x}{2} \geq 0$, i.e. $\mu \leq \frac{1}{2}$ and $1 - \frac{1}{2}b\Delta x > 0$, then

$$\begin{aligned} \left\|e^{n+1}\right\|_\infty &\leq \left(\mu + \frac{b\mu\Delta x}{2}\right)\|e^n\|_\infty + ((1-2\mu))\|e^n\|_\infty + \left(\mu - \frac{b\mu\Delta x}{2}\right)\|e^n\|_\infty + c\Delta t(\Delta t + (\Delta x)^2) \\ &= \|e^n\|_\infty + c\Delta t(\Delta t + (\Delta x)^2). \end{aligned}$$

Then

$$\begin{aligned} \|e^n\|_\infty &\leq \left\|e^{n-1}\right\|_\infty + c\Delta t(\Delta t + (\Delta x)^2) \\ &\leq \left\|e^{n-2}\right\|_\infty + c2\Delta t(\Delta t + (\Delta x)^2) \\ &\leq \quad \vdots \\ &\leq \left\|e^0\right\|_\infty + cn\Delta t(\Delta t + (\Delta x)^2) \\ &= ct^*(\Delta t + (\Delta x)^2). \end{aligned}$$

◀

B.2 Numerical Mathematics Preliminary Examination Aug. 2010

Problem B.9. *(Prelim Aug. 2010#1) Prove that $A \in \mathbb{C}^{m\times n}(m > n)$ and let $A = \hat{Q}\hat{R}$ be a reduced QR factorization.*

1. *Prove that A has rank n if and only if all the diagonal entries of $\hat{R}$ are non-zero.*
2. *Suppose* $\operatorname{rank}(A) = n$, *and define* $P = \hat{Q}\hat{Q}^*$. *Prove that* $\operatorname{range}(P) = \operatorname{range}(A)$.
3. *What type of matrix is P?*

Solution. 1. From the properties of reduced QR factorization, we know that $\hat{Q}$ has orthonormal columns, therefore $\det(\hat{Q}) = 1$ and $\hat{R}$ is upper triangular matrix, so $\det(\hat{R}) = \prod_{i=1}^n r_{ii}$. Then

$$\det(A) = \det(\hat{Q}\hat{R}) = \det(\hat{Q})\det(\hat{R}) = \prod_{i=1}^n r_{ii}.$$

Therefore, A has rank n if and only if all the diagonal entries of $\hat{R}$ are non-zero.

2. (a) $\operatorname{range}(A) \subseteq \operatorname{range}(P)$: Let $y \in \operatorname{range}(A)$, that is to say there exists a $x \in \mathbb{C}^n$ s.t. $Ax = y$. Then by reduced QR factorization we have $y = \hat{Q}\hat{R}x$. then

$$Py = P\hat{Q}\hat{R}x = \hat{Q}\hat{Q}^*\hat{Q}\hat{R}x = \hat{Q}\hat{R}x = Ax = y.$$

therefore $y \in \operatorname{range}(P)$.

(b) $\operatorname{range}(P) \subseteq \operatorname{range}(A)$: Let $v \in \operatorname{range}(P)$, that is to say there exists a $v \in \mathbb{C}^n$, s.t. $v = Pv = \hat{Q}\hat{Q}^*v$.

Claim B.1.

$$\hat{Q}\hat{Q}^* = A\left(A^*A\right)^{-1}A^*.$$

Proof.

$$\begin{aligned} A\left(A^*A\right)^{-1}A^* &= \hat{Q}\hat{R}\left(\hat{R}^*\hat{Q}^*\hat{Q}\hat{R}\right)^{-1}\hat{R}^*\hat{Q}^* \\ &= \hat{Q}\hat{R}\left(\hat{R}^*\hat{R}\right)^{-1}\hat{R}^*\hat{Q}^* \\ &= \hat{Q}\hat{R}\hat{R}^{-1}\left(\hat{R}^*\right)^{-1}\hat{R}^*\hat{Q}^* \\ &= \hat{Q}\hat{Q}^*. \end{aligned}$$

◀

Therefore by the claim, we have

$$v = Pv = \hat{Q}\hat{Q}^* v = A\left(A^*A\right)^{-1} A^* v = A\Big(\left(A^*A\right)^{-1} A^* v\Big) = Ax.$$

where $x = \left(A^*A\right)^{-1} A^* v$. Hence $v \in \text{range}(A)$.

3. P is an orthogonal projector.

◀

Problem B.10. *(Prelim Aug. 2010#4) Prove that $A \in \mathbb{R}^{n\times n}$ is SPD if and only if it has a Cholesky factorization.*

Solution. 1. Since A is SPD, so it has LU factorization, and $L = U$, i.e.

$$A = LU = U^T U.$$

Therefore, it has a Cholesky factorization.

2. if A has Cholesky factorization, i.e $A = U^T U$, then

$$x^T Ax = x^T U^T Ux = (Ux)^T Ux.$$

Let $y = Ux$, then we have

$$x^T Ax = (Ux)^T Ux = y^T y = y_1^2 + y_2^2 + \cdots + y_n^2 \geq 0,$$

with equality only when y = 0, i.e. x=0 (since U is non-singular). Hence A is SPD.

◀

Problem B.11. *(Prelim Aug. 2010#8) Consider the Crank-Nicolson scheme*

$$u_j^{n+1} = u_j^n + \frac{\mu}{2}\left(u_{j-1}^{n+1} - 2u_j^{n+1} + u_{j+1}^{n+1} + u_{j-1}^n - 2u_j^n + u_{j+1}^n\right)$$

for approximating the solution to the heat equation $\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$ on the intervals $0 \leq x \leq 1$ and $0 \leq t \leq t^*$ with the boundary conditions $u(0,t) = u(1,t) = 0$.

1. *Show that the scheme may be written in the form $\mathbf{u}^{n+1} = A\mathbf{u}^n$, where $A \in \mathbb{R}^{m\times m}_{sym}$(the space of $m \times m$ symmetric matrices) and*

$$\|Ax\|_2 \leq \|x\|_2,$$

for any $\mathbf{x} \in \mathbb{R}^m$, regardless of the value of μ.

2. *Show that*

$$\|Ax\|_\infty \leq \|x\|_\infty,$$

for any $\mathbf{x} \in \mathbb{R}^m$, provided $\mu \leq 1$.(In other words, the scheme may only be conditionally stable in the max norm.)

Solution. 1. the scheme

$$u_j^{n+1} = u_j^n + \frac{\mu}{2}\left(u_{j-1}^{n+1} - 2u_j^{n+1} + u_{j+1}^{n+1} + u_{j-1}^n - 2u_j^n + u_{j+1}^n\right)$$

can be rewritten as

$$-\frac{\mu}{2}u_{j-1}^{n+1} + (1+\mu)u_j^{n+1} - \frac{\mu}{2}u_{j+1}^{n+1} = \frac{\mu}{2}u_{j-1}^n + (1-\mu)u_j^n + \frac{\mu}{2}u_{j+1}^n.$$

By using the boundary, we have

$$C\mathbf{u}^{n+1} = B\mathbf{u}^n$$

where

$$C = \begin{bmatrix} 1+\mu & -\frac{\mu}{2} & & & \\ -\frac{\mu}{2} & 1+\mu & -\frac{\mu}{2} & & \\ & \ddots & \ddots & \ddots & \\ & & -\frac{\mu}{2} & 1+\mu & -\frac{\mu}{2} \\ & & & -\frac{\mu}{2} & 1+\mu \end{bmatrix}, B = \begin{bmatrix} 1-\mu & \frac{\mu}{2} & & & \\ \frac{\mu}{2} & 1-\mu & \frac{\mu}{2} & & \\ & \ddots & \ddots & \ddots & \\ & & \frac{\mu}{2} & 1-\mu & \frac{\mu}{2} \\ & & & \frac{\mu}{2} & 1-\mu \end{bmatrix},$$

$$\mathbf{u}^{n+1} = \begin{bmatrix} u_1^{n+1} \\ u_2^{n+1} \\ \vdots \\ u_m^{n+1} \end{bmatrix} \text{ and } \mathbf{u}^n = \begin{bmatrix} u_1^n \\ u_2^n \\ \vdots \\ u_m^n \end{bmatrix}.$$

So, the scheme may be written in the form $\mathbf{u}^{n+1} = A\mathbf{u}^n$, where $A = C^{-1}B$. By using the Fourier transform, i.e.

$$u_j^{n+1} = g(\xi)u_j^n, \;\; u_j^n = e^{ij\Delta x\xi},$$

then we have

$$-\frac{\mu}{2}g(\xi)u_{j-1}^n + (1+\mu)g(\xi)u_j^n - \frac{\mu}{2}g(\xi)u_{j+1}^n = \frac{\mu}{2}u_{j-1}^n + (1-\mu)u_j^n + \frac{\mu}{2}u_{j+1}^n.$$

And then

$$-\frac{\mu}{2}g(\xi)e^{i(j-1)\Delta x\xi} + (1+\mu)g(\xi)e^{ij\Delta x\xi} - \frac{\mu}{2}g(\xi)e^{i(j+1)\Delta x\xi} = \frac{\mu}{2}e^{i(j-1)\Delta x\xi} + (1-\mu)e^{ij\Delta x\xi} + \frac{\mu}{2}e^{i(j+1)\Delta x\xi},$$

i.e.

$$g(\xi)\left(-\frac{\mu}{2}e^{-i\Delta x\xi} + (1+\mu) - \frac{\mu}{2}e^{i\Delta x\xi}\right)e^{j\Delta x\xi} = \left(\frac{\mu}{2}e^{-i\Delta x\xi} + (1-\mu) + \frac{\mu}{2}e^{i\Delta x\xi}\right)e^{j\Delta x\xi},$$

i.e.

$$g(\xi)\left(1+\mu-\mu\cos(\Delta x\xi)\right) = 1-\mu+\mu\cos(\Delta x\xi).$$

therefore,

$$g(\xi) = \frac{1-\mu+\mu\cos(\Delta x\xi)}{1+\mu-\mu\cos(\Delta x\xi)}.$$

hence

$$g(\xi) = \frac{1+\frac{1}{2}z}{1-\frac{1}{2}z}, z = 2\frac{\Delta t}{\Delta x^2}(cos(\Delta x\xi)-1).$$

Moreover, $\left|g(\xi)\right| < 1$, therefore, $\rho(A) < 1$.

$$\|Ax\|_2 \le \|A\|_2 \|x\|_2 = \rho(A)\|x\|_2 \le \|x\|_2.$$

2. the scheme

$$u_j^{n+1} = u_j^n + \frac{\mu}{2}\left(u_{j-1}^{n+1} - 2u_j^{n+1} + u_{j+1}^{n+1} + u_{j-1}^n - 2u_j^n + u_{j+1}^n\right)$$

can be rewritten as

$$(1+\mu)u_j^{n+1} = \frac{\mu}{2}u_{j-1}^{n+1} + \frac{\mu}{2}u_{j+1}^{n+1} + \frac{\mu}{2}u_{j-1}^n + (1-\mu)u_j^n + \frac{\mu}{2}u_{j+1}^n.$$

then

$$\left|1+\mu\right|\left|u_j^{n+1}\right| \le \left|\frac{\mu}{2}\right|\left|u_{j-1}^{n+1}\right| + \left|\frac{\mu}{2}\right|\left|u_{j+1}^{n+1}\right| + \left|\frac{\mu}{2}\right|\left|u_{j-1}^n\right| + \left|(1-\mu)\right|\left|u_j^n\right| + \left|\frac{\mu}{2}\right|\left|u_{j+1}^n\right|.$$

Therefore

$$(1+\mu)\left\|u_j^{n+1}\right\|_\infty \le \frac{\mu}{2}\left\|u_{j-1}^{n+1}\right\|_\infty + \frac{\mu}{2}\left\|u_{j+1}^{n+1}\right\|_\infty + \frac{\mu}{2}\left\|u_{j-1}^n\right\|_\infty + \left|(1-\mu)\right|\left\|u_j^n\right\|_\infty + \frac{\mu}{2}\left\|u_{j+1}^n\right\|_\infty.$$

i.e.

$$(1+\mu)\left\|\mathbf{u}^{n+1}\right\|_\infty \le \frac{\mu}{2}\left\|\mathbf{u}^{n+1}\right\|_\infty + \frac{\mu}{2}\left\|\mathbf{u}^{n+1}\right\|_\infty + \frac{\mu}{2}\|\mathbf{u}^n\|_\infty + \left|(1-\mu)\right|\|\mathbf{u}^n\|_\infty + \frac{\mu}{2}\|\mathbf{u}^n\|_\infty.$$

if $\mu \le 1$, then

$$\left\|\mathbf{u}^{n+1}\right\|_\infty \le \|\mathbf{u}^n\|_\infty,$$

i.e.

$$\|A\mathbf{u}^n\|_\infty \le \|\mathbf{u}^n\|_\infty.$$

◀

Problem B.12. *(Prelim Aug. 2010#9) Consider the Lax-Wendroff scheme*

$$u_j^{n+1} = u_j^n + \frac{a^2(\Delta t)^2}{2(\Delta x)^2}\left(u_{j-1}^n - 2u_j^n + u_{j+1}^n\right) - \frac{a\Delta t}{2\Delta x}\left(u_{j+1}^n - u_{j-1}^n\right),$$

for the approximating the solution of the Cauchy problem for the advection equation

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0, a > 0.$$

Use Von Neumann's Method to show that the Lax-Wendroff scheme is stable provided the CFL condition

$$\frac{a\Delta t}{\Delta x} \le 1.$$

is enforced.

Solution. By using the Fourier transform, i.e.

$$u_j^{n+1} = g(\xi)u_j^n, \quad u_j^n = e^{ij\Delta x\xi},$$

then we have

$$g(\xi)u_j^n = u_j^n + \frac{a^2(\Delta t)^2}{2(\Delta x)^2}\left(u_{j-1}^n - 2u_j^n + u_{j+1}^n\right) - \frac{a\Delta t}{2\Delta x}\left(u_{j+1}^n - u_{j-1}^n\right).$$

And then

$$g(\xi)e^{ij\Delta x\xi} = e^{ij\Delta x\xi} + \frac{a^2(\Delta t)^2}{2(\Delta x)^2}\left(e^{i(j-1)\Delta x\xi} - 2e^{ij\Delta x\xi} + e^{i(j+1)\Delta x\xi}\right) - \frac{a\Delta t}{2\Delta x}\left(e^{i(j+1)\Delta x\xi} - e^{i(j-1)\Delta x\xi}\right).$$

Therefore

$$\begin{aligned} g(\xi) &= 1 + \frac{a^2(\Delta t)^2}{2(\Delta x)^2}\left(e^{-i\Delta x\xi} - 2 + e^{i\Delta x\xi}\right) - \frac{a\Delta t}{2\Delta x}\left(e^{i\Delta x\xi} - e^{-i\Delta x\xi}\right) \\ &= 1 + \frac{a^2(\Delta t)^2}{2(\Delta x)^2}\left(2\cos(\Delta x\xi) - 2\right) - \frac{a\Delta t}{2\Delta x}\left(2i\sin(\Delta x\xi)\right) \\ &= 1 + \frac{a^2(\Delta t)^2}{(\Delta x)^2}\left(\cos(\Delta x\xi) - 1\right) - \frac{a\Delta t}{\Delta x}\left(i\sin(\Delta x\xi)\right). \end{aligned}$$

Let $\mu = \frac{a\Delta t}{\Delta x}$, then

$$g(\xi) = 1 + \mu^2\left(\cos(\Delta x\xi) - 1\right) - \mu\left(i\sin(\Delta x\xi)\right).$$

If $\left|g(\xi)\right| < 1$, then the scheme is stable, i,e

$$\left(1 + \mu^2\left(\cos(\Delta x\xi) - 1\right)\right)^2 + \left(\mu\sin(\Delta x\xi)\right)^2 < 1.$$

i.e.

$$1 + 2\mu^2\left(\cos(\Delta x\xi) - 1\right) + \mu^4\left(\cos(\Delta x\xi) - 1\right)^2 + \mu^2\sin(\Delta x\xi)^2 < 1.$$

i.e.

$$\mu^2\left(\sin(\Delta x\xi)^2 + 2\cos(\Delta x\xi) - 2\right) + \mu^4\left(\cos(\Delta x\xi) - 1\right)^2 < 0.$$

i.e.

$$\mu^2\left(1 - \cos(\Delta x\xi)^2 + 2\cos(\Delta x\xi) - 2\right) + \mu^4\left(\cos(\Delta x\xi) - 1\right)^2 < 0.$$

i.e

$$\mu^2(\cos(\Delta x\xi) - 1)^2 - (\cos(\Delta x\xi) - 1)^2 < 0,$$

$$(\mu^2 - 1)(\cos(\Delta x\xi) - 1)^2 < 0,$$

then we get $\mu < 1$. The above process is invertible, therefore, we prove the result. ◄

B.3 Numerical Mathematics Preliminary Examination Jan. 2009

B.4 Numerical Mathematics Preliminary Examination Jan. 2008

Problem B.13. *(Prelim Jan. 2008#8) Let $\Omega \subset \mathbb{R}^2$ be a bounded domain with a smooth boundary. Consider a 2-D poisson-like equation*

$$\begin{cases} -\Delta u + 3u &= x^2y^2, \text{ in } \Omega, \\ u &= 0, \text{ on } \partial\Omega. \end{cases}$$

1. *Write the corresponding Ritz and Galerkin variational problems.*
2. *Prove that the Galerkin method has a unique solution u_h and the following estimate is valid*

$$\|u - u_h\|_{H^1} \leq C \inf_{v_h \in V_h} \|u - v_h\|_{H^1},$$

with C independent of h, where V_h denotes a finite element subspace of $H^1(\Omega)$ consisting of continuous piecewise polynomials of degree $k \geq 1$.

Solution. 1. For this pure Dirichlet Problem, the test functional space $v \in H_0^1$. Multiple the test function on the both sides of the original function and integral on Ω, we get

$$-\int_\Omega \Delta u v dx + \int_\Omega uv dx = \int_\Omega xyv dx.$$

Integration by part yields

$$\int_\Omega \nabla u \nabla v dx + \int_\Omega uv dx = \int_\Omega xyv dx.$$

Let

$$a(u,v) = \int_\Omega \nabla u \nabla v dx + \int_\Omega uv dx, \quad f(v) = \int_\Omega xyv dx.$$

Then, the

(a) Ritz variational problem is: find $u_h \in H_0^1$, such that

$$J(u_h) = \min \frac{1}{2} a(u_h, u_h) - f(u_h).$$

(b) Galerkin variational problem is: find $u_h \in H_0^1$, such that

$$a(u_h, u_h) = f(u_h).$$

2. Next, we will use Lax-Milgram to prove the uniqueness.

(a)

$$\begin{aligned} \left|a(u,v)\right| &\leq \int_\Omega |\nabla u \nabla v| dx + \int_\Omega |uv| dx \\ &\leq \|\nabla u\|_{L^2(\Omega)} \|\nabla v\|_{L^2(\Omega)} + \|u\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} \\ &\leq \|\nabla u\|_{L^2(\Omega)} \|\nabla v\|_{L^2(\Omega)} + C \|\nabla u\|_{L^2(\Omega)} \|\nabla v\|_{L^2(\Omega)} \\ &\leq C \|\nabla u\|_{L^2(\Omega)} \|\nabla v\|_{L^2(\Omega)} \\ &\leq C \|u\|_{H^1(\Omega)} \|v\|_{H^1(\Omega)} \end{aligned}$$

(b)

$$a(u,u) = \int_\Omega (\nabla u)^2 dx + \int_\Omega u^2 dx$$

So,

$$\begin{aligned} \left|a(u,u)\right| &= \int_\Omega |\nabla u|^2 dx + \int_\Omega |u|^2 dx \\ &= \|\nabla u\|^2_{L^2(\Omega)} + \|u\|^2_{L^2(\Omega)} \\ &= \|u\|^2_{H^1(\Omega)}. \end{aligned}$$

(c)

$$\begin{aligned} \left|f(v)\right| &\le \int_\Omega \left|xyv\right| dx \\ &\le \max|xy| \int_\Omega |v| dx \\ &\le C\left(\int_\Omega 1^2 dx\right)^{1/2} \left(\int_\Omega |v|^2 dx\right)^{1/2} \\ &\le C\|v\|_{L^2(\Omega)} \le C\|v\|_{H^1(\Omega)}. \end{aligned}$$

by Lax-Milgram theorem, we get that e Galerkin method has a unique solution u_h. Moreover,

$$a(v_h, v_h) = f(v_h).$$

And from the weak formula, we have

$$a(u, v_h) = f(v_h).$$

then we get the Galerkin Orthogonal (GO)

$$a(u - u_h, v_h) = 0.$$

Then by coercivity

$$\begin{aligned} \|u - u_h\|^2_{H^1(\Omega)} &\le \left|a(u - u_h, u - u_h)\right| \\ &= \left|a(u - u_h, u - v_h) + a(u - u_h, v_h - u_h)\right| \\ &= \left|a(u - u_h, u - v_h)\right| \\ &\le \|u - u_h\|_{H^1(\Omega)} \|u - v_h\|_{H^1(\Omega)}. \end{aligned}$$

Therefore,

$$\|u - u_h\|_{H^1} \le C \inf_{v_h \in V_h} \|u - v_h\|_{H^1},$$

◀

C Project 1 MATH571

COMPUTATIONAL ASSIGNMENT # 1

MATH 571

1. Instability of Gram–Schmidt

The purpose of the first part of your assignment is to investigate the instability of the classical Gram–Schmidt orthogonalization process. Lecture 9 in BT is somewhat related to this and could be a good source of inspiration.

1.– Write a piece of code that implements the classical Gram–Schmidt process, see Algorithm 7.1 in BT. Ideally, this should be implemented in the form of a QR factorization, that is, given a matrix $A \in \mathbb{R}^{m\times n}$ your method should return two matrices $Q \in \mathbb{R}^{m\times n}$ and $R \in \mathbb{R}^{n\times n}$, where the matrix Q has (or at least should have) orthonormal columns and $A = QR$.

2.– With the help of the developed piece of code, test the algorithm on a matrix $A = \mathbb{R}^{20\times 10}$ with:

- entries uniformly distributed over the interval $[0, 1]$.
- entries given by
$$a_{i,j} = \left(\frac{2i-21}{19}\right)^{j-1}.$$
- entries given by
$$a_{i,j} = \frac{1}{i+j-1},$$
this is the so-called *Hilbert matrix*.

3.– For each one of these cases compute $Q^\star Q$. Since Q, in theory, has orthonormal columns what should you get? What do you actually get?

4.– Implement the modified Gram-Schmidt process (Algorithm 8.1 in BT) and repeat steps 1.—3. What do you observe?

2. Linear Least Squares

The purpose of the second part of your assignment is to observe the so-called *Runge's phenomenon* and try to mitigate it using least squares. Lecture 11 on BT might give some hints on how to proceed. Consider the function

$$f(x) = \frac{1}{1+25x^2},$$

on the interval $[-1, 1]$. Do the following:

1.– Choose $N \in \mathbb{N}$ (not too large ≈ 10 should suffice) and on an equally spaced grid of points construct a polynomial that interpolates f. In other words, given the grid of points

$$x_i = -1 + \frac{2i}{N}, \quad i = \overline{0, N},$$

Date: Due October 16, 2013.

you must find a polynomial p_N of degree N such that

$$p_N(x_i) = f(x_i), \qquad i = \overline{0, N}.$$

2.– Even though f and p_N coincide at the nodes, how do they compare on the whole interval? You can, for instance plot them or look at their values on a grid that consists of $2N$ points.

3.– We are going to, instead of interpolating, construct a least squares fit for f. In other words, we choose $n \in \mathbb{N}$, $n < N$, and construct a polynomial q_n of degree n such that

$$\sum_{i=0}^{N} |q_n(x_i) - f(x_i)|^2$$

is minimal.

4.– If our least squares polynomial is defined as $q_n(x) = \sum_{j=0}^{n} Q_j x^j$, then the minimality conditions lead to the overdetermined system

$$\mathbf{Aq} = \mathbf{y}, \quad \mathbf{A}_{i,j} = x_i^{j-1}, \quad \mathbf{Q}_j = Q_j, \quad \mathbf{y}_i = f(x_i), \tag{1}$$

which, since all the points x_i are different has full rank (*Can you prove this?*). This means that the least squares solution can be found, for instance, using the QR algorithm which you developed on the first part of the assignment. This gives you the coefficients of the polynomial.

5.– How does q_n and f compare? Keeping N fixed, vary n and try to find an empirical relation for the n (in terms of N) which optimizes the least squares fit.

Remark. Equation (1) is also the system of equations you obtain when trying to compute the interpolating polynomial of point 1.–. In this case, however, the system will be square. You can still use the QR algorithm to solve this system.

MATH 571: Coding Assignment #1

Due on Wednesday, October 16, 2013

TTH 12:40pm

Wenqiang Feng

Wenqiang Feng MATH 571 (TTH 12:40pm): Coding Assignment #1

Contents

Wenqiang Feng MATH 571 (TTH 12:40pm): Coding Assignment #1

Problem 1

1. See Listing 3.
2. See Listing 2.
3. We should get the Identity square matrices. But we did not get the actual Identity square matrices through the Gram-Schmidt Algorithm. For case 1-2, we only get the matrices which $diag(Q^*Q) = \{\overbrace{1, \cdots, 1}^{n}\}$ and the other elements approximate to 0 in the sense of $C \times 10^{-16} \sim 10^{-17}$. For case 3, Classical Gram-Schmidt Algorithm is not stable for case 3, since some elements of matrix Q^*Q do not approximate to 0, then the matrix Q^*Q is not diagonal any more.
4. For case 1-2, we also did not get the actual Identity square matrices by using the Modified Gram-Schmidt Algorithm. We only get the matrices which $diag(Q^*Q) = \{\overbrace{1, \cdots, 1}^{n}\}$ and the other elements approximate to 0 in the sense of $C \times 10^{-17} \sim 10^{-18}$. For case 3, the Modified Gram-Schmidt Algorithm works well for case 3, we get the matrix which $diag(Q^*Q) = \{\overbrace{1, \cdots, 1}^{n}\}$ and the other elements approximate to 0 in the sense of $C \times 10^{-8} \sim 10^{-13}$. So, Modified Gram-Schmidt Algorithm is more stable than the Classical one.

Listing 1 shows the main function for problem1.

Listing 1: Main Function of Problem1

```
%Main function
clc
clear all
m=20;n=10;
fun1=@(i,j) ((2*i-21)/19)^(j-1);
fun2=@(i,j) 1/(i+j);
A1=rand(m,n);
A2=matrix_gen(m,n,fun1);
A3=matrix_gen(m,n,fun2);
% Test  for the random case 1
[CQ1,CR1]=gschmidt(A1)
[MQ1,MR1]=mgschmidt(A1)
q11=CQ1'*CQ1
q12=MQ1'*MQ1
% Test for case 2
[CQ2,CR2]=gschmidt(A2)
[MQ2,MR2]=mgschmidt(A2)
q21=CQ2'*CQ2
q22=MQ2'*MQ2
% Test for case 3
[CQ3,CR3]=gschmidt(A3)
[MQ3,MR3]=mgschmidt(A3)
q31=CQ3'*CQ3
q32=MQ3'*MQ3
```

Listing 2 shows the matrices generating function.

Listing 2: Matrices Generating Function

```
function A=matrix_gen(m,n,fun)
A=zeros(m,n);
for i=1:m
    for j=1:n
        A(i,j)=fun(i,j);
    end
end
```

Listing 3 shows Classical Gram-Schmidt Algorithm.

Listing 3: Classical Gram-Schmidt Algorithm

```
function [Q,R]=gschmidt(V)
% gschmidt:  classical Gram-Schmidt algorithm
%
% USAGE
%        gschmidt(V)
%
% INPUT
%      V: V is an m by n matrix of full rank m<=n
%
% OUTPUT
%     Q: an m-by-n matrix with orthonormal columns
%     R: an n-by-n upper triangular matrix
%
% AUTHOR
%    Wenqiang Feng
%    Department of Mathematics
%    University of Tennessee at Knoxville
%    E-mail: wfeng@math.utk.edu
%    Date:   9/14/2013

[m,n]=size(V);
Q=zeros(m,n);
R=zeros(n);
R(1,1)=norm(V(:,1));
Q(:,1)=V(:,1)/R(1,1);
for k=2:n
    R(1:k-1,k)=Q(:,1:k-1)'*V(:,k);
    Q(:,k)=V(:,k)-Q(:,1:k-1)*R(1:k-1,k);
    R(k,k)=norm(Q(:,k));
        if R(k,k) == 0
             break;
        end
    Q(:,k)=Q(:,k)/R(k,k);
end
```

Listing 4 shows Modified Gram-Schmidt Algorithm.

Listing 4: Modified Gram-Schmidt Algorithm

```
function [Q,R]=mgschmidt(V)
```

```
% mgschmidt:   Modified Gram-Schmidt algorithm
%
% USAGE
%        mgschmidt(V)
%
% INPUT
%      V: V is an m by n matrix of full rank m<=n
%
% OUTPUT
%     Q: an m-by-n matrix with orthonormal columns
%     R: an n-by-n upper triangular matrix
%
% AUTHOR
%    Wenqiang Feng
%    Department of Mathematics
%    University of Tennessee at Knoxville
%    E-mail: wfeng@math.utk.edu
%    Date:   9/14/2013

[m,n]=size(V);
Q=zeros(m,n);
R=zeros(n);

for k=1:n
    R(k,k)=norm(V(:,k));
        if R(k,k) == 0
             break;
        end
    Q(:,k)=V(:,k)/R(k,k);
        for j=k+1:n
            R(k,j)=Q(:, k)'* V(:,j);
            V(:, j) = V(:, j)-R(k, j) * Q(:,k);
        end
end
```

Problem 2

1. I Chose $N = 10$ and I got the polynomial p_{10} is as follow:

$$\begin{aligned} P_10 \quad &= -220.941742081448x^{10} + 7.38961566181029e^{-13}x^9 + 494.909502262444x^8 \\ &-1.27934383856085e^{-12}x^7 - 381.433823529411x^6 + 5.56308212237901e^{-13}x^5 \\ &+123.359728506787x^4 - 1.16016030941682e^{-14}x^3 - 16.8552036199095x^2 \\ &-5.86232968237562e^{-15}x + 1.00000000000000 \end{aligned}$$

2. See Figure 1.

3. See Listing 6.

4. Since A is Vandermonde Matrix and all the points x_i are different, then $det(A) \neq 0$. Therefore A has full rank.

5. I varied N from 3 to 15. For every fixed N, I varied n form 1 to N. Then I got the following table (Table.1). From table (Table.1), we can get that $n \approx 2\sqrt{N+1}$, where the N is the number of the partition.

$N \setminus h$	1	2	3	4	5	6	7	8	$\cdots$
3	0.23	$3.96 \cdot 10^{-17}$	$5.55 \cdot 10^{-17}$						
4	0.82	0.56	0.56	$5.10 \cdot 10^{-17}$					
5	0.50	0.28	0.28	$9.04 \cdot 10^{-16}$	$9.32 \cdot 10^{-16}$				
6	0.84	0.62	0.62	0.43	0.43	$8.02 \cdot 10^{-15}$			
7	0.71	0.46	0.46	0.25	0.25	$3.32 \cdot 10^{-15}$	$3.96 \cdot 10^{-15}$		
8	0.89	0.64	0.64	0.45	0.45	0.30	0.30	$1.39 \cdot 10^{-14}$	
$\vdots$									

Table 1: The L^2 norm of the Least squares polynomial fit

Fix $N = 10$, vary n(Figure 2-Figure 11).

Listing 5 shows main function of problem2.1.

Listing 5: Main Function of Problem2.1

```
% Main function of A2
clc
clear all
N=10;
n=N;
fun= @(x) 1./(1+25*x.^2);
x=-1:2/N:1;
y=fun(x);

x1=-1:2/(2*N):1;
a = polyfit(x,y,n);
p = polyval(a,x1)
plot(x,y,'o',x1,p,'-')

for m=1:10
least_squares(x, y, m)
end
```

Listing 6 shows Polynomial Least Squares Fitting Algorithm.

Listing 6: Polynomial Least Squares Fitting Algorithm

```
%Main function for pro#2.5
clc
clear all
for N=3:15
j=1;
for n=1:N%3:N;
fun= @(x) 1./(1+25*x.^2);
```

```
x=-1:2/N:1;
b=fun(x);

A=MatrixGen(x,n);
cof=GSsolver(A,b);
q=0;
     for i=1:n+1
        q=q+cof(i)*(x.^(i-1));
     end
error(j)=norm(q-b);
j=j+1;
error
end
end


function A=MatrixGen(x,n)
m=size(x,2);
A=zeros(m,n+1);
for i=1:m
    for j=1:n+1
        A(i,j)=x(i).^(j-1);
    end
 end


function x=GSsolver(A,b)
   [Q,R]=mgschmidt(A);
    x= R\(Q'*b');
```

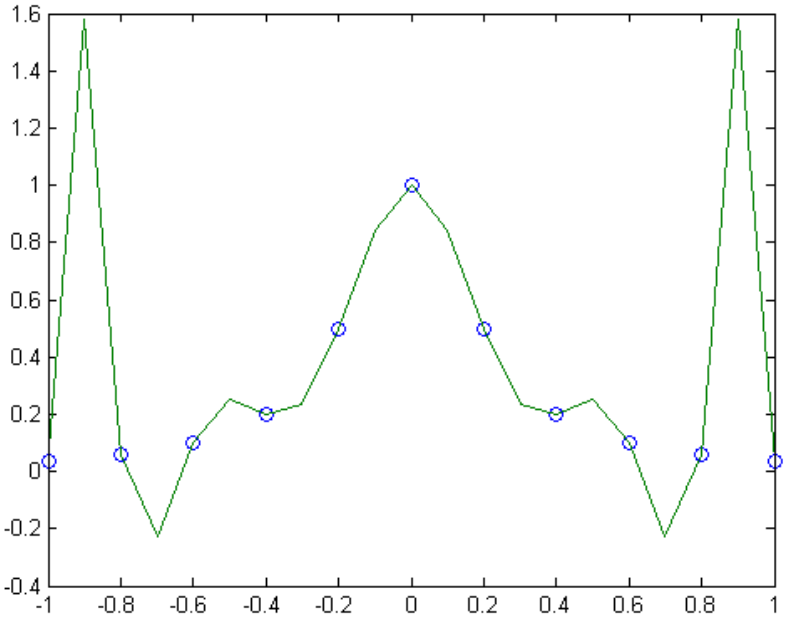

Figure 1: Runge's phenomenon of Polynomial interpolation with 2N points.

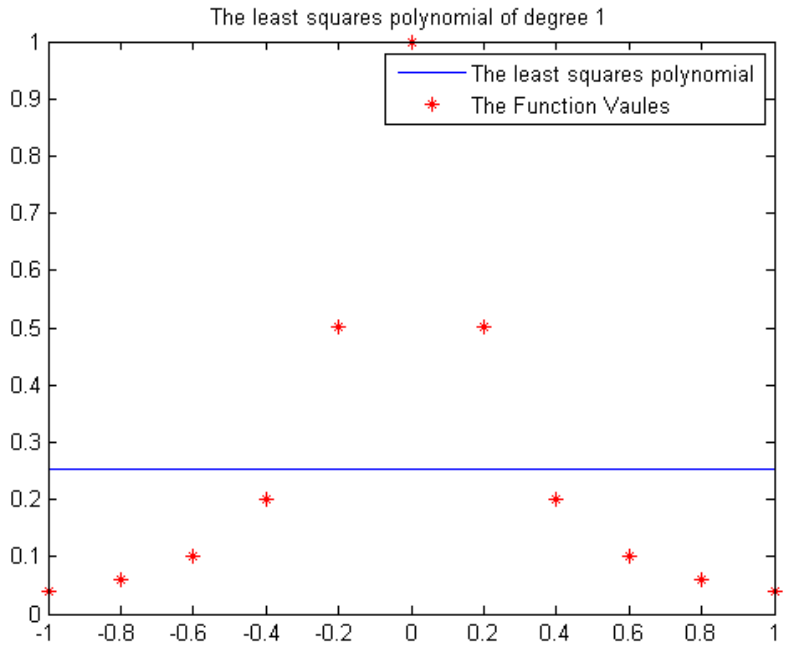

Figure 2: Least Square polynomial of degree=1, N=10.

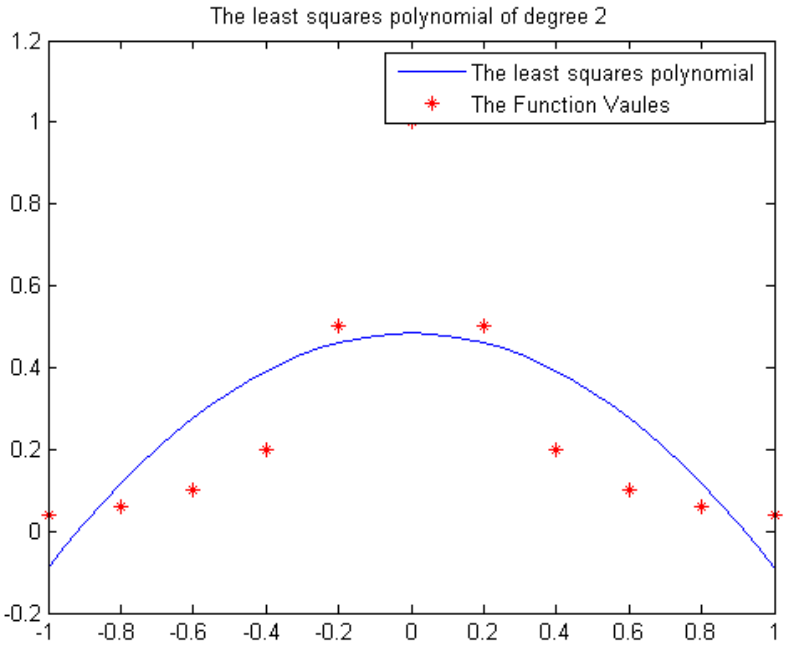

Figure 3: Least Square polynomial of degree=2, N=10.

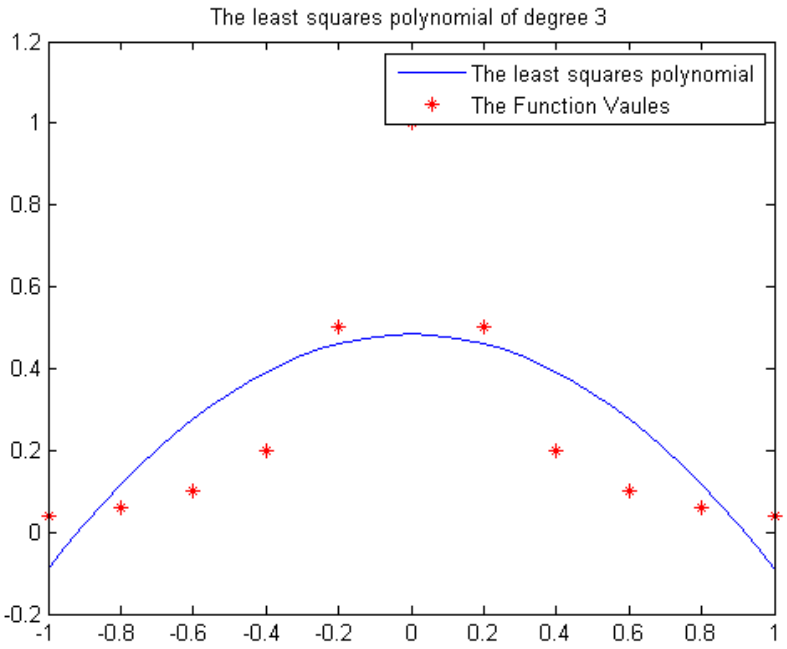

Figure 4: Least Square polynomial of degree=3, N=10.

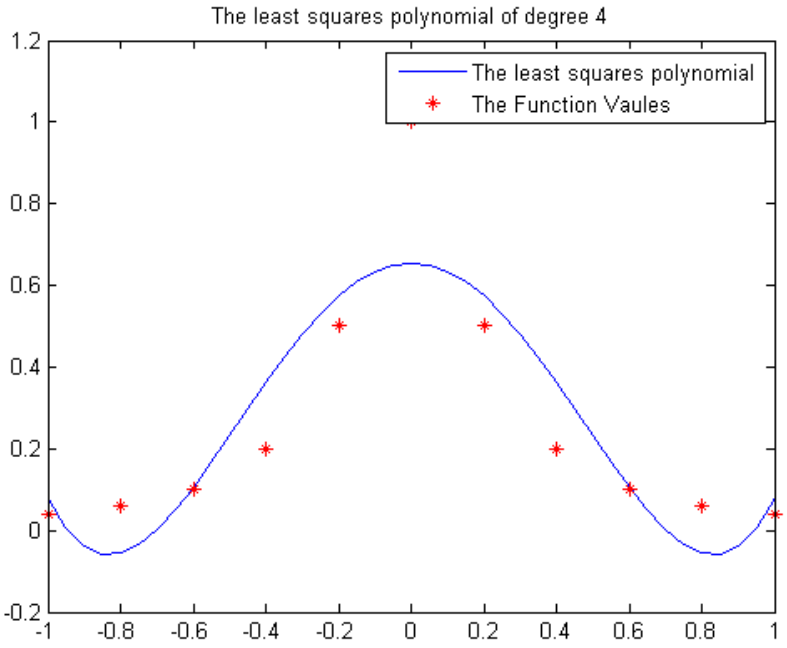

Figure 5: Least Square polynomial of degree=4, N=10.

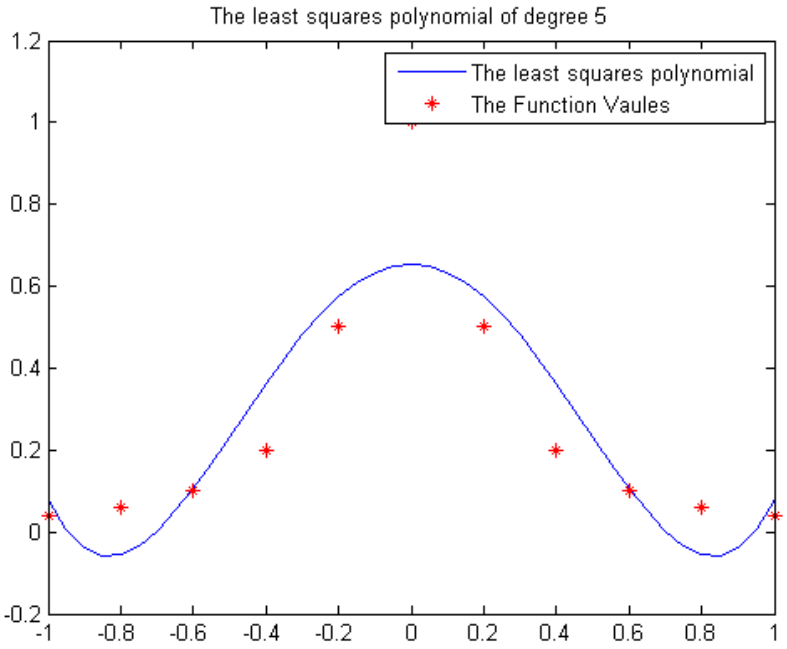

Figure 6: Least Square polynomial of degree=5, N=10.

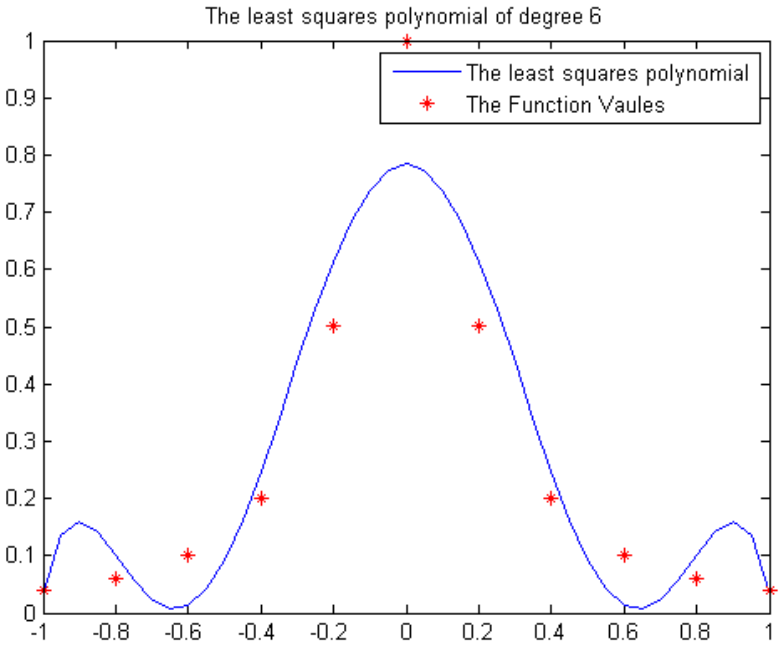

Figure 7: Least Square polynomial of degree=6, N=10.

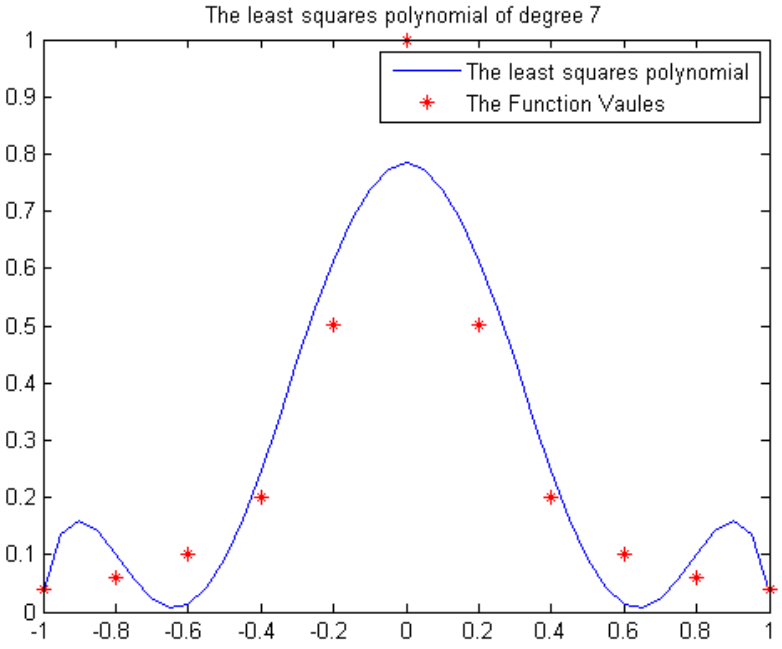

Figure 8: Least Square polynomial of degree=7, N=10.

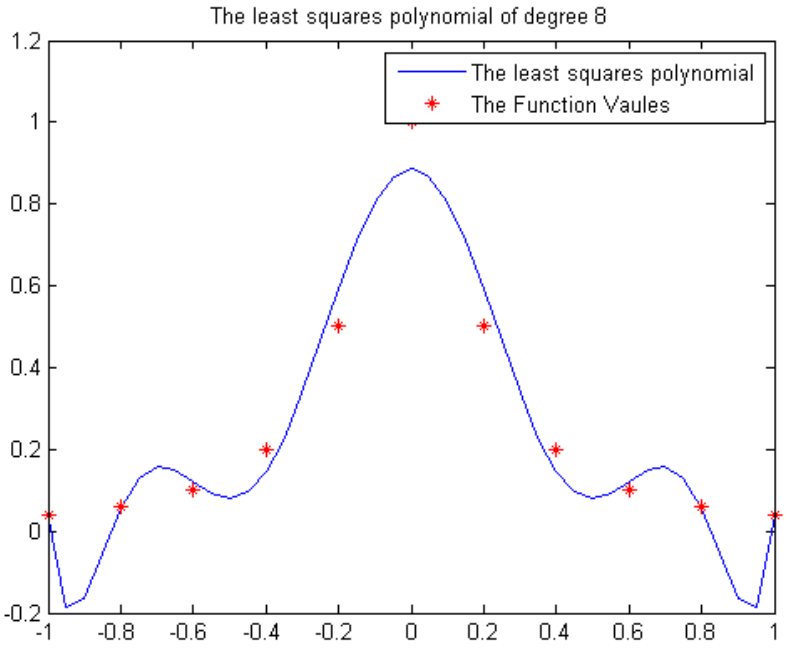

Figure 9: Least Square polynomial of degree=8, N=10.

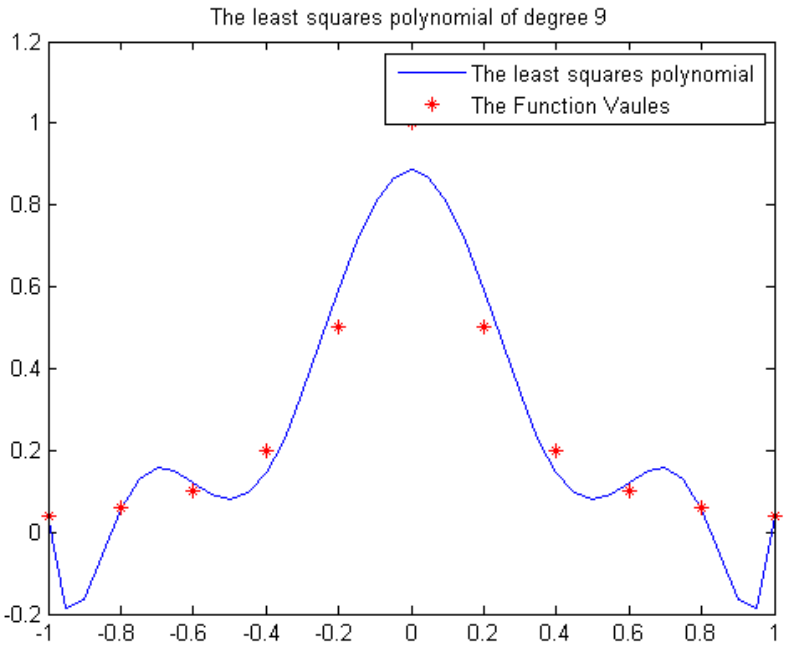

Figure 10: Least Square polynomial of degree=9, N=10.

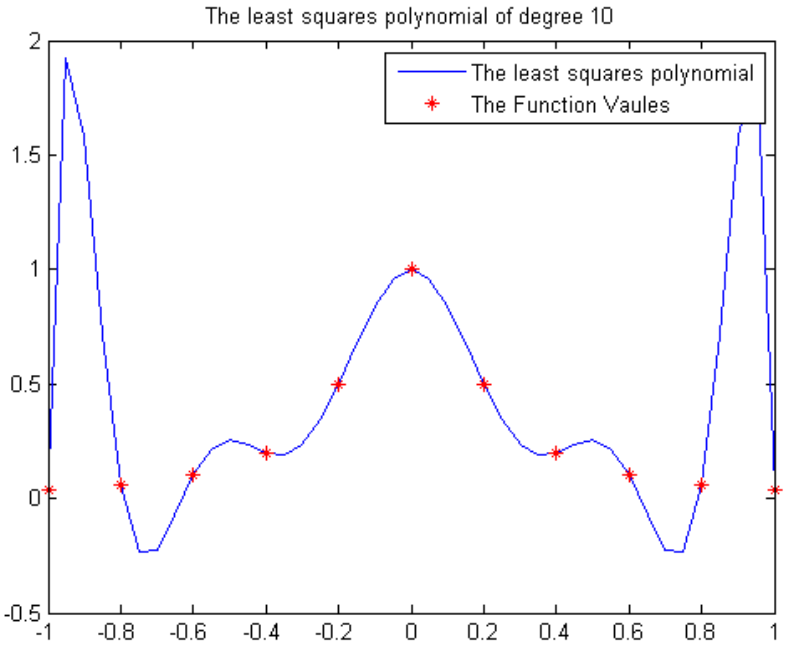

Figure 11: Least Square polynomial of degree=10, N=10.

D Project 2 MATH571

COMPUTATIONAL ASSIGNMENT # 2

MATH 571

1. Convergence of Classical Schemes

The purpose of this part of your assignment is to investigate the convergence properties of classical iterative schemes. To do so, develop:

1. A piece of code $[\tilde{x}, K] = \mathsf{Jacobi}(M, f, \epsilon)$ that implements the Jacobi method.
2. A piece of code $[\tilde{x}, K] = \mathsf{SOR}(M, f, \omega, \epsilon)$ that implements the SOR method. Notice that the number ω should be an input parameter[1].

Your implementations should take as input a square matrix $M \in \mathbb{R}^{N\times N}$, a right hand side vector $f \in \mathbb{R}^N$ and a tolerance $\epsilon > 0$. The output should be a vector $\tilde{x} \in \mathbb{R}^N$ — an approximate solution to $Mx = f$ and an integer K — the number of iterations.[2]

For $n \in \mathbb{N}$ set $N = 2n - 1$ and consider the following matrices:

- The nonsymmetric matrix $A \in \mathbb{R}^{N\times N}$:

$$A_{i,i} = 3,\ i = \overline{1, N}, \qquad A_{i,i+1} = -1,\ i = \overline{1, N-1}, \qquad A_{i,i-n} = -1,\ i = \overline{n+1, N}.$$

- The tridiagonal matrix $J \in \mathbb{R}^{N\times N}$:

$$J_{1,1} = 1 = -J_{1,2}, \quad J_{i,i} = 2 + \frac{1}{N^2},\ J_{i,i+1} = J_{i,i-1} = -1,\ i = \overline{2, N-1}, \quad J_{N,N} = 1 = -J_{N,N-1}.$$

- The tridiagonal matrix $S \in \mathbb{R}^{N\times N}$:

$$S_{i,i} = 3,\ i = \overline{1, N}, \qquad S_{i,i+1} = -1,\ i = \overline{1, N-1}, \qquad S_{i,i-1} = -1,\ i = \overline{2, N}.$$

For different values of $n \in \{2, \dots, 50\}$ and for each $M \in \{A, J, S\}$, choose a vector $x \in \mathbb{R}^N$ and define $f_M = Mx$.

i) Run $\mathsf{Jacobi}(M, f_M, \epsilon)$ and record the number of iterations. How does the number of iterations depend on N?
ii) Run $\mathsf{SOR}(M, f_M, 1, \epsilon)$. How does the number of iterations depend on N?
iii) Try to find the optimal value of ω, that is the one for which the number of iterations is minimal.
iv) How does the number of iterations between $\mathsf{Jacobi}(M, f_M, \epsilon)$, $\mathsf{SOR}(M, f_M, 1, \epsilon)$ and $\mathsf{SOR}(M, f_M, \omega)$ with an optimal ω compare? What can you conclude?

2. The Method of Alternating Directions

In this section we will study the Alternating Directions Implicit (ADI) method. Given $A \in \mathbb{R}^{N\times N}$, $A = A^\star > 0$ and $f \in \mathbb{R}^N$ we wish to solve $Ax = f$. Assume that we have the following *splitting* of the matrix A:

$$A = A_1 + A_2, \qquad A_i = A_i^\star > 0,\ i = 1, 2, \qquad A_1 A_2 = A_2 A_1.$$

Date: Due November 26, 2013.

[1] Recall that for $\omega = 1$ we obtain the Gauß–Seidel method so you obtain two methods for one here ;-)

[2] As stopping criterion you can use either

$$\|x^{k+1} - x^k\| < \epsilon,$$

or, since we are just trying to learn,

$$\|x^{k+1} - x\| < \epsilon,$$

where x is the exact solution.

Then, we propose the following scheme

$$(I+\tau A_1)\frac{x^{k+1/2}-x^k}{\tau}+Ax^k=f, \tag{1}$$

$$(I+\tau A_2)\frac{x^{k+1}-x^{k+1/2}}{\tau}+Ax^{k+1/2}=f. \tag{2}$$

1. Write a piece of code $[\tilde{x},K]=\mathsf{ADI}(A,f,\epsilon,A_1,A_2,\tau)$ that implements the ADI scheme described above. As before, the input should be a matrix $A\in\mathbb{R}^N$, a right hand side $f\in\mathbb{R}^N$ and a tolerance $\epsilon>0$. In addition, the scheme should take parameters $A_1,A_2\in\mathbb{R}^N$ and $\tau>0$.

   Notice that, in general, we need to invert $(I+\tau A_i)$. In practice these matrices are chosen so that these inversions are easy.
2. Let $n\in\{4,\dots,50\}$ and set $N=n^2$. Define the matrices $\Lambda,\Sigma\in\mathbb{R}^{N\times N}$ as follows:

```
for i = 1,n
  for j = 1,n
    I = i + n(j-1);
    Λ[I,I] = Σ[I,I] = 3;
    if i < n
      Λ[I,I+1] = -1;
    endif
    if i > 1
      Λ[I,I-1] = -1;
    endif
    if j < n
      Σ[I,I+n] = -1;
    endif
    if j > 1
      Σ[I,I-n] = -1;
    endif
  endfor
endfor
```

3. Set $A=\Lambda+\Sigma$.
4. Are the matrices Λ and Σ SPD? Do they commute?
5. Choose $x\in\mathbb{R}^N$ and set $f=Ax$. Run $\mathsf{ADI}(A,f,\epsilon,\Lambda,\Sigma,\tau)$ for different values of τ. Which one seems to be the optimal one?

   The following are not obligatory but can be used as extra credit:
6. Write an expression for x^{k+1} in terms of x^k only. *Hint:* Try adding and subtracting (1) and (2). What do you get?
7. From this expression find the equation that controls the error $e^k=x-x^k$.
8. Assume that $(A_1A_2x,x)\geq 0$, show that in this case $[x,y]=(A_1A_2x,y)$ is an inner product. If that is the case we will denote $\|x\|_B=[x,x]^{1/2}$.
9. Under this assumption we will show convergence of the ADI scheme. To do so:
   - Take the inner product of the equation that controls the error with $e^{k+1}+e^k$.
   - Add over $k=\overline{1,K}$. We should obtain

$$\|e^{K+1}\|_2^2+\tau\sum_{k=1}^{K}\|e^{k+1}+e^k\|_A^2+2\tau^2\|e^{K+1}\|_B^2=\|e^0\|_2^2+2\tau^2\|e^0\|_B^2.$$

   - From this it follows that, for every $\tau>0$, $\frac{1}{2}(x^{k+1}+x^k)\to x$. How?

MATH 571: Computational Assignment #2

Due on Tuesday, November 26, 2013

TTH 12:40pm

Wenqiang Feng

Wenqiang Feng MATH 571 (TTH 12:40pm): Computational Assignment #2

Contents

Wenqiang Feng MATH 571 (TTH 12:40pm): Computational Assignment #2

Let $Ndim$ to be the Dimension of the matrix and $Niter$ to be the iterative numbers. In the whole report, b was generated by Ax, where x is a corresponding vector and x's entries are random numbers between 0 and 10. The initial iterative values of x are given by $\vec{0}$.

Problem 1

1. Listing 1 shows the implement of Jacobi Method.
2. Listing 2 shows the implement of SOR Method.
3. The numerical results:
   (a) From the records of the iterative number, I got the following results:
   For case (2), the Jacobi Method is not convergence, because it has a big Condition Number. For case (1) and case (3), if $Ndim$ is small, roughly speaking, $Ndim \leq 10 - 20$, then the $Ndim$ and $Niter$ have the roughly relationship $Niter = log(Ndim + C)$, when $Ndim$ is large, the $Niter$ is not depent on the $Ndim$ (see Figure (1)).
   (b) When $\omega = 1$, the SOR Method degenerates to the Gauss-seidel Method. For Gauss-seidel Method, I get the similar results as Jacobi Method (see Figure (2)). But, the Gauss-Seidel Method is more stable than Jacobi Method and case (3) is more stable than case (1) (see Figure (1) and Figure (2)).

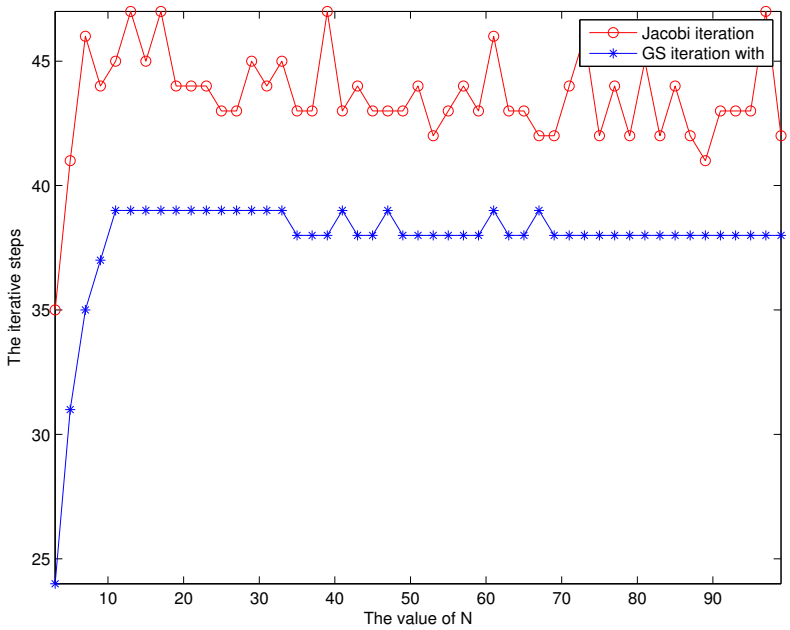

Figure 1: The relationship between $Ndim$ and $Niter$ for case(1)

   (c) The optimal w
      i. For case (1), the optimal w is around 1, but this optimal w is not optimal for all (see Figure (3) and Figure (4));

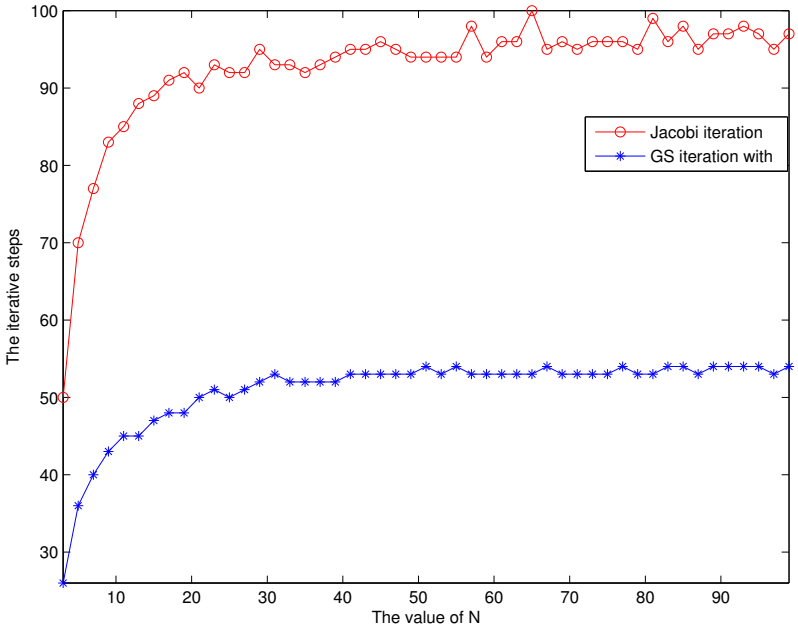

Figure 2: The relationship between $Ndim$ and $Niter$ for case (3)

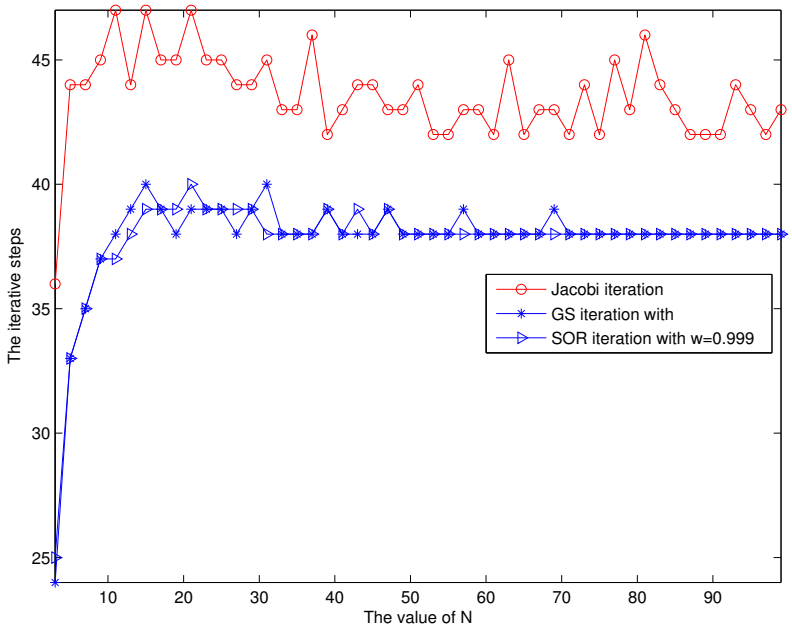

Figure 3: The relationship between $Ndim$ and $Niter$ for case(1)

ii. For case (2), In general, the SOR Method is not convergence, but SOR is convergence for some small $Ndim$;

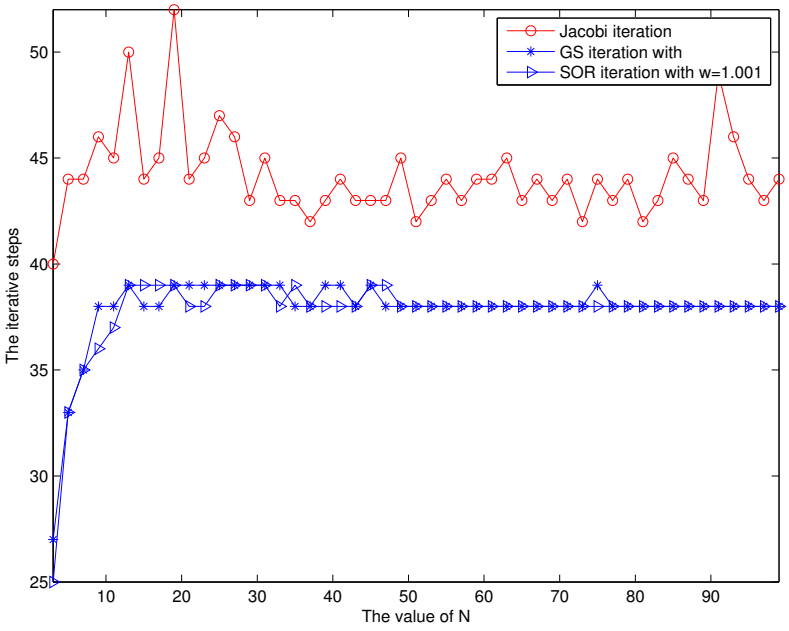

Figure 4: The relationship between $Ndim$ and $Niter$ for case(1)

iii. For case(3), the optimal w is around 1.14; This numerical result is same as the theoretical result. Let D = $diag(diag(A))$; $E = A - D$; $T = D \backslash E$,

$$w_{opt} = \frac{2}{\sqrt{1-\rho(T)^2}} \approx 1.14.$$

Where, the $\rho(T)$ is the spectral radius of T (see Figure (5)).

(d) In general, for the convergence case, $Niter_{Jacobi} > Niter_{Gauss-Sediel} > Niter_{SOR_{opt}}$. I conclude that SOR_{opt} is more efficient than $Gauss - Sediel$ and $Gauss - Sediel$ is more efficient than $Jacobi$ for convergence case (see Figure (5)).

Listing 1: Jacobi Method

```
function [x iter]=jacobi(A,b,x,tol,max_iter)
% jacobi:  Solve the linear system with Jacobi iterative algorithm
%
% USAGE
%        jacobi(A,b,x0,tol)
%
% INPUT
%      A: N by N LHS coefficients matrix
%      b: N by 1 RHS vector
%      x: Initial guess
%      tol: The stop tolerance
%      max_iter: maxmum iterative steps
%
% OUTPUT
%     x: The solutions
```

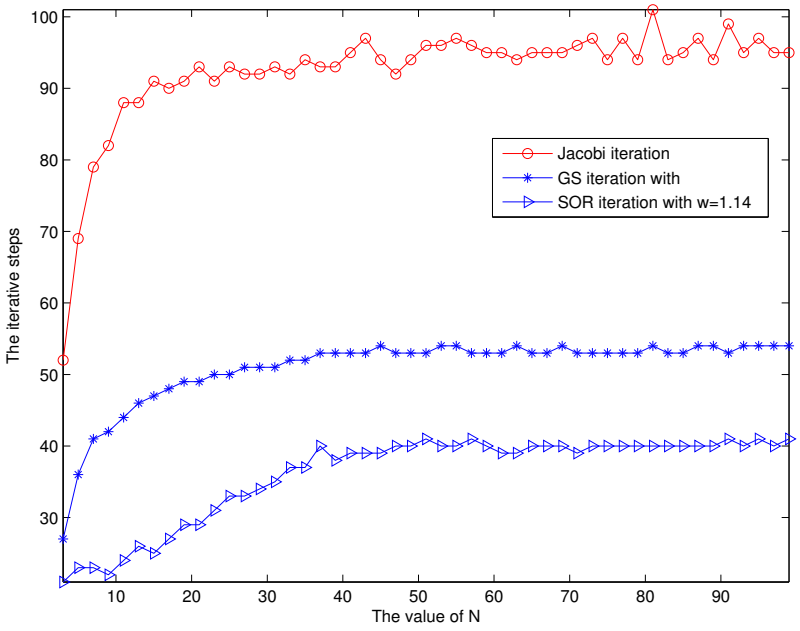

Figure 5: The relationship between $Ndim$ and $Niter$ for case(3)

```
%     iter: iterative steps
%
% AUTHOR
%    Wenqiang Feng
%    Department of Mathematics
%    University of Tennessee at Knoxville
%    E-mail: wfeng@math.utk.edu
%    Date:   11/13/2013
n=size(A,1);

% Set default parameters
if (nargin<3), x=zeros(n,1);tol=1e-16;max_iter=500;end;
%Initial some parameters
error=norm(b - A*x);
iter=0 ;
%split the matrix for Jacobi interative method
D = diag(diag(A));
E=D-A;

while (error>tol&&iter<max_iter)
    x1=x;
    x= D\(E*x+b);
    error=norm(x-x1);
    iter=iter+1;
end
```

Listing 3: SOR Method

```
function [x iter]=sor(A,b,w,x,tol,max_iter)
% jacobi:  Solve the linear system with SOR iterative algorithm
%
% USAGE
%        jacobi(A,b,epsilon,x0,tol,max_iter)
%
% INPUT
%      A: N by N LHS coefficients matrix
%      b: N by 1 RHS vector
%      w: Relaxation parameter
%      x: Initial guess
%      tol: The stop tolerance
%      max_iter: maxmum iterative steps
%
% OUTPUT
%     x: The solutions
%     iter: iterative steps
%
% AUTHOR
%    Wenqiang Feng
%    Department of Mathematics
%    University of Tennessee at Knoxville
%    E-mail: wfeng@math.utk.edu
%    Date:   11/13/2013

n=size(A,1);
% Set default parameters
if (nargin<4), x=zeros(n,1);tol=1e-16;max_iter=500;end;
%Initial some parameters
error=norm(b - A*x)/norm( b );
iter=0 ;
%split the matrix for Jacobi interative method
     D=diag(diag( A ));
     b = w * b;
     M =  w * tril( A, -1 ) + D;
     N = -w * triu( A,  1 ) + ( 1.0 - w ) * D;

while (error>tol&&iter<max_iter)
    x1=x;
    x= M\(N*x+b);
    error=norm(x-x1)/norm( x );
    iter=iter+1;
end
```

Problem 2

1. Listing 3 shows the implement of ADI Method.

2. Yes, The Σ and Λ are the SPD matrices. Moreover, they are commute, since $\Sigma\Lambda = \Lambda\Sigma$.

3. The optimal τ for the ADI method:
   The optimal τ for the ADI method is same as the $SSOR$ and SOR method. Let D = $diag(diag(A))$; $E = A - D$; $T = D\backslash E$,
   $$\tau_{opt} = \frac{2}{\sqrt{1-\rho(T)^2}}.$$
   Where, the $\rho(T)$ is the spectral radius of T.

4. The expression of x^{k+1}:
   By adding and subtracting scheme (1) and scheme (2), we get that
   $$(I+\tau A_1)(I+\tau A_2)x^{k+1} - (I-\tau A_1)(I-\tau A_2)x^k = 2\tau f. \tag{1}$$

5. The expression of the error's control:
   $$(I+\tau A_1)(I+\tau A_2)e^{k+1} = (I-\tau A_1)(I-\tau A_2)e^k. \tag{2}$$

6. Now, I will show $[x,y] = (A_1A_2x, y)$ is an inner product, i.e, I will show the $||x||_B^2 = [x,x]$ satisfies parallelogram law:
   It's easy to show that the B-norm $||x||_B^2 = [x,x]$ satisfies the parallelogram law,
   $$\begin{aligned} ||x+y||_B^2 + ||x-y||_B^2 &= (A_1A_2(x+y), x+y) + (A_1A_2(x-y), x-y) \\ &= (A_1A_2x, x) + (A_1A_2x, y) + (A_1A_2y, x) + (A_1A_2y, y) \\ &\quad +(A_1A_2x, x) - (A_1A_2x, y) - (A_1A_2y, x) + (A_1A_2y, y) \\ &= 2(||x||_B^2 + ||y||_B^2). \end{aligned}$$
   So, The norm space can induce a inner product, so $[x,y] = (A_1A_2x, y)$ is a inner product.

7. Take inner product (2) with $e^{k+1} + e^k$, we get,
   $$\left((I+\tau A_1)(I+\tau A_2)e^{k+1}, e^{k+1}+e^k\right) = \left((I-\tau A_1)(I-\tau A_2)e^k, e^{k+1}+e^k\right). \tag{3}$$
   By using the distribution law, we get
   $$\begin{aligned} &\left(e^{k+1}, e^{k+1}\right) + \tau\left(Ae^{k+1}, e^{k+1}\right) + \tau^2\left(A_1A_2e^{k+1}, e^{k+1}\right) && (4)\\ &+\left(e^{k+1}, e^k\right) + \tau\left(Ae^{k+1}, e^k\right) + \tau^2\left(A_1A_2e^{k+1}, e^k\right) && (5)\\ &= \left(e^k, e^{k+1}\right) - \tau\left(Ae^k, e^{k+1}\right) + \tau^2\left(A_1A_2e^k, e^{k+1}\right) && (6)\\ &+\left(e^k, e^k\right) - \tau\left(Ae^k, e^k\right) + \tau^2\left(A_1A_2e^k, e^k\right). && (7) \end{aligned}$$
   Since, $A_1A_2 = A_2A_1$, so $\left(A_1A_2e^{k+1}, e^k\right) = \left(A_1A_2e^k, e^{k+1}\right)$. Therefore, (4) reduces to
   $$\begin{aligned} &\left(e^{k+1}, e^{k+1}\right) + \tau\left(A(e^{k+1}+e^k), e^{k+1}+e^k\right) + \tau^2\left(A_1A_2e^{k+1}, e^{k+1}\right) && (8)\\ =\ &\left(e^k, e^k\right) + \tau^2\left(A_1A_2e^k, e^k\right). && (9) \end{aligned}$$
   Therefore,
   $$||e^{k+1}||_2^2 + \tau||e^{k+1}+e^k||_A^2 + \tau^2||e^{k+1}||_B^2 = ||e^k||_2^2 + \tau^2||e^k||_B^2. \tag{10}$$

Summing over k from 0 to K, we get

$$||e^{K+1}||_2^2 + \tau \sum_{k=0}^{K} ||e^{k+1} + e^k||_A^2 + \tau^2 ||e^{K+1}||_B^2 = ||e^0||_2^2 + \tau^2 ||e^0||_B^2. \tag{11}$$

Therefore, from (11), we get $||e^{k+1} + e^k||_A^2 \to 0 \ \forall \tau > 0$. So $\frac{1}{2}(x^{k+1} + x^k) \to x$ with respect to $||\cdot||_A$.

Listing 3: ADI Method

```
function [x iter]=adi(A,b,A1,A2,tau,x,tol,max_iter)
% jacobi:  Solve the linear system with ADI algorithm
%
% USAGE
%        adi(A,b,A1,A2,tau,x,tol,max_iter)
%
% INPUT
%      A: N by N LHS coefficients matrix
%      b: N by 1 RHS vector
%      A1: The decomposition of A: A=A1+A2 and A1*A2=A2*A1
%      A2: The decomposition of A: A=A1+A2 and A1*A2=A2*A1
%      x: Initial guess
%      tol: The stop tolerance
%      max_iter: maxmum iterative steps
%
% OUTPUT
%     x: The solutions
%     iter: iterative steps
%
% AUTHOR
%    Wenqiang Feng
%    Department of Mathematics
%    University of Tennessee at Knoxville
%    E-mail: wfeng@math.utk.edu
%    Date:   11/13/2013
n=size(A,1);

% Set default parameters
if (nargin<6), x=zeros(n,1);tol=1e-16;max_iter=300;end;

%Initial some parameters
error=norm(b - A*x);
iter=0 ;
I=eye(n);

while (error>tol&&iter<max_iter)
    x1=x;
    x=(tau*I+A1)\((tau*I-A2)*x+b); % the first half step
    x=(tau*I+A2)\((tau*I-A1)*x+b); % the second half step
    error=norm(x-x1);
    iter=iter+1;
end
```

E Midterm examination 572

MATH 572: Exam problem 4-5

Due on July 15, 2014

TTH 12:40pm

Wenqiang Feng

Wenqiang Feng MATH 572 (TTH 12:40pm): Exam problem 4-5

Contents

Problem 1

Given the equation

$$\begin{cases} -u'' + u = f, & \text{in } \Omega \\ -u'(0) = u'(1) = 0, & \text{on } \partial\Omega \end{cases} \tag{1}$$

devise a finite difference scheme for this problem that results in a tridiagonal matrix. The scheme must be consistent of order $\mathcal{O}(2)$ in the $C(\bar{\Omega}_h)$ norm and you should prove this.

Proof: I consider the following uniform partition (Figure. 1) of the interval $(0, 1)$ with $N+1$ points. For the Neumann Boundary, we introduce two ghost point x_{-1} and x_{N+1}.

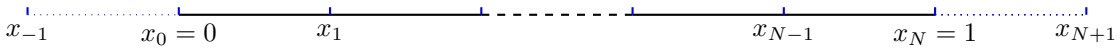

Figure 1: One dimension's partition

The second order scheme of (1) is as following

$$\begin{cases} -\frac{U_{i+1}-2U_i+U_{i-1}}{h^2} + U_i & = F_i, \quad \forall i = 0, \cdots, N, \\ -\frac{U_1 - U_{-1}}{2h} & = 0, \\ \frac{U_{N+1}-U_{N-1}}{2h} & = 0. \end{cases} \tag{2}$$

From the homogeneous Neumann boundary condition, we know that $U_1 = U_{-1}$ and $U_{N+1} = U_{N-1}$. Therefore

1. for $i = 0$, from the scheme,
$$-\frac{1}{h^2}U_{-1} + \frac{2}{h^2}U_0 - \frac{1}{h^2}U_1 + U_0 = (1 + \frac{2}{h^2})U_0 - \frac{2}{h^2}U_1 = F_0$$
2. for $i = 1, \cdots, N-1$, we get
$$-\frac{1}{h^2}U_{i-1} + \frac{2}{h^2}U_i - \frac{1}{h^2}U_{i+1} + U_i = -\frac{1}{h^2}U_{i-1} + (1 + \frac{2}{h^2})U_i - \frac{1}{h^2}U_{i+1} = F_i.$$
3. for $i = N$
$$-\frac{1}{h^2}U_{N-1} + \frac{2}{h^2}U_N - \frac{1}{h^2}U_{N+1} + U_N = (1 + \frac{2}{h^2})U_N - \frac{2}{h^2}U_{N-1} = F_N.$$

So the algebraic system is

$$AU = F,$$

where

$$A = \begin{pmatrix} 1+\frac{2}{h^2} & -\frac{2}{h^2} & & & \\ -\frac{1}{h^2} & 1+\frac{2}{h^2} & -\frac{1}{h^2} & & \\ & \ddots & \ddots & \ddots & \\ & & -\frac{1}{h^2} & 1+\frac{2}{h^2} & -\frac{1}{h^2} \\ & & & -\frac{2}{h^2} & 1+\frac{2}{h^2} \end{pmatrix}, U = \begin{pmatrix} U_0 \\ U_1 \\ \vdots \\ U_{N-1} \\ U_N \end{pmatrix} F = \begin{pmatrix} F_0 \\ F_1 \\ \vdots \\ F_{N-1} \\ F_N \end{pmatrix}.$$

Next, I will show this scheme is of order $\mathcal{O}(2)$. From the Taylor expansion, we know

$$U_{i+1} = u(x_{i+1}) = u(x_i) + hu'(x_i) + \frac{h^2}{2}u''(x_i) + \frac{h^3}{2}u^{(3)}(x_i) + \mathcal{O}(h^4)$$

$$U_{i-1} = u(x_{i-1}) = u(x_i) - hu'(x_i) + \frac{h^2}{2}u''(x_i) - \frac{h^3}{2}u^{(3)}(x_i) + \mathcal{O}(h^4).$$

Therefore,

$$-\frac{U_{i+1}-2U_i+U_{i-1}}{h^2}=-\frac{u(x_{i+1})-2u(x_i)+u(x_{i-1})}{h^2}=-u''(x_i)+\mathcal{O}(h^2).$$

Therefore, the scheme (2) is of order $\mathcal{O}(h^2)$.

Problem 2

Let $A=tridiag\{a_i,b_i,c_i\}_{i=1}^n\in\mathbb{R}^{n\times n}$ be a tridiagional matrix with the properties that

$$b_i>0,\quad a_i,c_i\le 0,\quad a_i+b_i+c_i=0.$$

Prove the following maximum principle: If $u\in\mathbb{R}^n$ is such that $(Au)_{i=2,\cdots,n-1}\le 0$, then $u_i\le\max\{u_1,u_n\}$.

Proof: Without loss generality, we assume $u_k, k=2,\cdots,n-1$ is the maximum value.

1. For $(Au)_{i=2,\cdots,n-1}<0$:
   I will use the method of contradiction to prove this case. Since $(Au)_{i=2,\cdots,n-1}<0$, so
   $$a_ku_{k-1}+b_ku_k+c_ku_{k+1}<0.$$
   Since $a_k+c_k=-b_k$ and $a_k<0, c_k<0$, so
   $$a_ku_{k-1}-(a_k+c_k)u_k+c_ku_{k+1}=a_k(u_{k-1}-u_k)+c_k(u_{k+1}-u_k)\ge 0.$$
   This is contradiction to $(Au)_{i=2,\cdots,n-1}<0$. Therefore, If $u\in\mathbb{R}^n$ is such that $(Au)_{i=2,\cdots,n-1}<0$, then $u_i\le\max\{u_1,u_n\}$.
2. For $(Au)_{i=2,\cdots,n-1}=0$:
   Since $(Au)_{i=2,\cdots,n-1}=0$, so
   $$a_ku_{k-1}+b_ku_k+c_ku_{k+1}=0.$$
   Since $a_k+c_k=-b_k$, so
   $$a_ku_{k-1}-(a_k+c_k)u_k+c_ku_{k+1}=a_k(u_{k-1}-u_k)+c_k(u_{k+1}-u_k)=0.$$
   And $a_k<0, c_k<0, u_{k-1}-u_k\le 0, u_{k+1}-u_k\le 0$, so $u_{k-1}=u_k=u_{k+1}$, that is to say, u_{k-1} and u_{k+1} is also the maximum points. Bu using the same argument again, we get $u_{k-2}=u_{k-1}=u_k=u_{k+1}=u_{k+2}$. Repeating the process, we get
   $$u_1=u_2=\cdots=u_{n-1}=u_n.$$
   Therefore, If $u\in\mathbb{R}^n$ is such that $(Au)_{i=2,\cdots,n-1}=0$, then $u_i\le\max\{u_1,u_n\}$

Problem 3

Prove the following discrete Poincaré inequality: Let $\Omega=(0,1)$ and Ω_h be a uniform grid of size h. If $Y\in\mathcal{U}_h$ is a mesh function on Ω_h such that $Y(0)=0$, then there is a constant C, independent of Y and h, for which

$$\|Y\|_{2,h}\le C\left\|\bar{\delta}Y\right\|_{2,h}.$$

Proof: I consider the following uniform partition (Figure. 2) of the interval $(0,1)$ with N points.

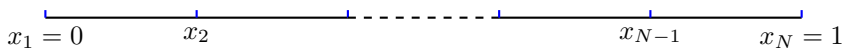

Figure 2: One dimension's uniform partition

Since the discrete 2-norm is defined as follows

$$\|v\|_{2,h}^2 = h^d \sum_{i=1}^{N} |v_i|^2,$$

where d is dimension. So, we have

$$\|v\|_{2,h}^2 = h\sum_{i=1}^{N} |v_i|^2, \quad \|\bar{\delta} v\|_{2,h}^2 = h\sum_{i=2}^{N} \left|\frac{v_{i-1} - v_i}{h}\right|^2.$$

Since $Y(0) = 0$, i.e. $Y_1 = 0$,

$$\sum_{i=2}^{N} Y_{i-1} - Y_i = Y_1 - Y_N = -Y_N.$$

Then,

$$\left|\sum_{i=2}^{N} Y_{i-1} - Y_i\right| = |Y_N|.$$

and

$$|Y_N| \le \sum_{i=2}^{N} |Y_{i-1} - Y_i| = \sum_{i=2}^{N} h\left|\frac{Y_{i-1} - Y_i}{h}\right| \le \left(\sum_{i=2}^{N} h^2\right)^{1/2} \left(\sum_{i=2}^{N} \left|\frac{Y_{i-1} - Y_i}{h}\right|^2\right)^{1/2}.$$

Therefore

$$\begin{aligned} |Y_K|^2 &\le \left(\sum_{i=2}^{K} h^2\right)\left(\sum_{i=2}^{K} \left|\frac{Y_{i-1} - Y_i}{h}\right|^2\right) \\ &= h^2(K-1)\sum_{i=2}^{K} \left|\frac{Y_{i-1} - Y_i}{h}\right|^2. \end{aligned}$$

1. When $K = 2$,

$$|Y_2|^2 \le h^2 \left|\frac{Y_1 - Y_2}{h}\right|^2.$$

2. When $K = 3$,

$$|Y_3|^2 \le 2h^2 \left(\left|\frac{Y_1 - Y_2}{h}\right|^2 + \left|\frac{Y_2 - Y_3}{h}\right|^2\right).$$

3. When $K = N$,

$$|Y_N|^2 \le (N-1)h^2 \left(\left|\frac{Y_1 - Y_2}{h}\right|^2 + \left|\frac{Y_2 - Y_3}{h}\right|^2 + \cdots + \left|\frac{Y_{N-1} - Y_N}{h}\right|^2\right).$$

Sum over $|Y_i|^2$ from 2 to N, we get

$$\sum_{i=2}^{N} |Y_i|^2 \leq \frac{N(N-1)}{2} h^2 \sum_{i=2}^{N} \left| \frac{Y_{i-1} - Y_i}{h} \right|^2 .$$

Since $Y_1 = 0$, so

$$\sum_{i=1}^{N} |Y_i|^2 \leq \frac{N(N-1)}{2} h^2 \sum_{i=2}^{N} \left| \frac{Y_{i-1} - Y_i}{h} \right|^2 .$$

And then

$$\frac{1}{(N-1)^2} \sum_{i=1}^{N} |Y_i|^2 \leq \frac{N}{2(N-1)} h^2 \sum_{i=2}^{N} \left| \frac{Y_{i-1} - Y_i}{h} \right|^2 = \left(\frac{1}{2} + \frac{1}{2(N-1)} \right) h^2 \sum_{i=2}^{N} \left| \frac{Y_{i-1} - Y_i}{h} \right|^2 .$$

Since $h = \frac{1}{N-1}$, so

$$h^2 \sum_{i=1}^{N} |Y_i|^2 \leq \left(\frac{1}{2} + \frac{1}{2(N-1)} \right) h^2 \sum_{i=2}^{N} \left| \frac{Y_{i-1} - Y_i}{h} \right|^2 .$$

then

$$h \sum_{i=1}^{N} |Y_i|^2 \leq \left(\frac{1}{2} + \frac{1}{2(N-1)} \right) h \sum_{i=2}^{N} \left| \frac{Y_{i-1} - Y_i}{h} \right|^2 .$$

i.e,

$$\|Y\|_{2,h}^2 \leq \left(\frac{1}{2} + \frac{1}{2(N-1)} \right) \left\| \bar{\delta} Y \right\|_{2,h}^2 .$$

since $N \geq 2$, so

$$\|Y\|_{2,h}^2 \leq \left\| \bar{\delta} Y \right\|_{2,h}^2 .$$

Hence,

$$\|Y\|_{2,h} \leq C \left\| \bar{\delta} Y \right\|_{2,h} .$$

F Project 1 MATH572

COMPUTATIONAL ASSIGNMENT # 1

MATH 572

Adaptive Solution of Ordinary Differential Equations

All the theorems about convergence that we have had in class state that, under certain conditions,

$$\lim_{h\to 0+} \max_n \|y(t_n) - y_n\| = 0.$$

While this is good and we should **not** use methods that do not satisfy this condition, this type of result is of little help in practice. In other words, we usually compute with a fixed h and, even if we know $y(t_n)$, we do not know the exact solution at the next time step and, thus, cannot assess how small the local error

$$e_{n+1} = y(t_{n+1}) - y_{n+1}$$

is. Here we will study two strategies to estimate this quantity. Your assignment will consist in implementing these two strategies and use them for the solution of a Cauchy problem

$$y' = f(t,y) \quad t \in (t_0, T), \qquad y(t_0) = y_0,$$

where

1. $f = y - t$, $(t_0, T) = (0, 10)$, $y_0 = 1 + \delta$, with $\delta \in \{0, 10^{-3}\}$.
2. $f = \lambda y + \sin t - \lambda \cos t$, $(t_0, T) = (0, 5)$, $y_0 = 0$, $\lambda \in \{0, \pm 5, \pm 10\}$.
3. $f = 1 - \frac{y}{t}$, $(t_0, T) = (2, 20)$, $y_0 = 2$.
4. The Fresnel integral is given by
$$\phi(t) = \int_0^t \sin(s^2) ds.$$
Set it as a Cauchy problem and generate a table of values on $[0, 10]$. If possible obtain a plot of the function.
5. The dilogarithm function
$$f(x) = -\int_0^x \frac{\ln(1-t)}{t} dt$$
on the interval $[-2, 0]$.

Step Bisection. The local error analysis that is usually carried out with the help of Taylor expansions yields, for a method of order s, that

$$\|e_{n+1}\| \le Ch^{s+1}.$$

The constant C here is independent of h but it might depend on the exact solution y and the current step t_n. To control the local error we will assume that C *does not change* as n changes. Let v denote the value of the approximate solution at t_{n+1} obtained by doing one step of length h from t_n. Let u be the approximate solution at t_{n+1} obtained by taking *two* steps of size $h/2$ from t_n. The important thing here is that both u and v are *computable*. By the assumption on C we have

$$y(t_{n+1}) = v + Ch^{s+1},$$
$$y(t_{n+1}) = u + 2C(h/2)^{s+1},$$

which implies

$$\|e_{n+1}\| \le Ch^{s+1} = \frac{\|u - v\|}{1 - 2^{-s}}.$$

Notice that the quantity on the right of this expression is completely computable. In a practical realization one can then monitor $\|u - v\|$ to make sure that it is below a prescribed tolerance. If it is not, the time step

Date: Due March 13, 2014.

can be reduced (halved) to improve the local truncation error. On the other hand, if this quantity is well below the prescribed tolerance, the time step can be doubled.

Implement this strategy for the fourth order ERK scheme

$$\begin{array}{c|cccc} 0 & & & & \\ \frac{1}{2} & \frac{1}{2} & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & & \\ 1 & 0 & 0 & 1 & \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array}$$

Adaptive Runge-Kutta-Fehlberg Method. The Runge-Kutta-Fehlberg method is an attempt at devising a procedure to automatically choose the step size. It consists of a fourth order and a fifth order method with cleverly chosen parameters so that they use the same nodes and, thus, the function evaluations are at the same points. The result is a fifth order method that has an estimate for the local error. The method computes two sequences $\{y_n\}$ and $\{\bar{y}_n\}$ of fifth and fourth order, respectively, by

$$\begin{array}{c|c} \mathbf{c} & \mathbf{A} \\ \hline & \mathbf{b}^\intercal \\ & \bar{\mathbf{b}}^\intercal \end{array} = \begin{array}{c|cccccc} 0 & & & & & & \\ \frac{1}{4} & \frac{1}{4} & & & & & \\ \frac{3}{8} & \frac{3}{32} & \frac{9}{32} & & & & \\ \frac{12}{13} & \frac{1932}{2197} & -\frac{7200}{2197} & \frac{7296}{2197} & & & \\ 1 & \frac{439}{216} & -8 & \frac{3680}{513} & -\frac{845}{4104} & & \\ \frac{1}{2} & -\frac{8}{27} & 2 & -\frac{3544}{2565} & \frac{1859}{4104} & -\frac{11}{40} & \\ \hline & \frac{16}{135} & 0 & \frac{6656}{12825} & \frac{28561}{56430} & -\frac{9}{50} & \frac{2}{55} \\ & \frac{25}{216} & 0 & \frac{1408}{2565} & \frac{2197}{4104} & -\frac{1}{5} & 0 \end{array}$$

The quantity

$$e_{n+1} = y_{n+1} - \bar{y}_{n+1} = h\sum_{i=1}^{6}(b_i - \bar{b}_i)f(t_n + c_i h, \xi_i)$$

can be used as an estimate of the local error. An algorithm to control the step size is based on the size of $\|y_{n+1} - \bar{y}_{n+1}\|$ which, in principle, is controlled by Ch^5.

Implement this scheme.

MATH 572: Computational Assignment #2

Due on Thurday, March 13, 2014

TTH 12:40pm

Wenqiang Feng

Wenqiang Feng MATH 572 (TTH 12:40pm): Computational Assignment #2

Contents

Wenqiang Feng MATH 572 (TTH 12:40pm): Computational Assignment #2

Adaptive Runge-Kutta Methods Formulas

In this project, we consider two adaptive Runge-Kutta Methods for the following initial-value ODE problem

$$\begin{cases} y'(t) = f(t,y) \\ y(t_0) = y_0, \end{cases} \tag{1}$$

The formula for the fourth order Runge-Kutta (4th RK) method can be read as following

$$\begin{cases} y(t_0) = y_0, \\ K_1 = hf(t_i, y_i) \\ K_2 = hf(t_i + \frac{h}{2}, y_i + \frac{K_1}{2}) \\ K_3 = hf(t_i + \frac{h}{2}, y_i + \frac{K_2}{2}) \\ K_4 = hf(t_i + h, y_i + K_3) \\ y_{i+1} = y_i + \frac{1}{6}(K_1 + K_2 + K_3 + K_4) \end{cases} \tag{2}$$

And the Adaptive Runge-Kutta-Fehlberg (RKF) Method can be wrote as

$$\begin{cases} y(t_0) = y_0, \\ K_1 = hf(t_i, y_i) \\ K_2 = hf(t_i + \frac{h}{4}, y_i + \frac{K_1}{4}) \\ K_3 = hf(t_i + \frac{3h}{8}, y_i + \frac{3}{32}K_1 + \frac{9}{32}K_2) \\ K_4 = hf(t_i + \frac{12h}{13}, y_i + \frac{1932}{2197}K_1 - \frac{7200}{2197}K_2 + \frac{7296}{2197}K_3) \\ K_5 = hf(t_i + h, y_i + \frac{439}{216}K_1 - 8K_2 + \frac{3680}{513}K_3 - \frac{845}{4104}K_4) \\ K_6 = hf(t_i + \frac{h}{2}, y_i - \frac{8}{27}K_1 + 2K_2 - \frac{3544}{2565}K_3 + \frac{1859}{4104}K_4 - \frac{11}{40}) \\ y_{i+1} = y_i + \frac{16}{135}K_1 + \frac{6656}{12825}K_3 + \frac{28561}{56430}K_4 - \frac{9}{50}K_5 + \frac{2}{55}K_6 \\ \tilde{y}_{i+1} = y_i + \frac{25}{216}K_1 + \frac{1408}{2656}K_3 + \frac{2197}{4104}K_4 - \frac{1}{5}K_5. \end{cases} \tag{3}$$

The error

$$E = \frac{1}{h}|y_{i+1} - \tilde{y}_{i+1}| \tag{4}$$

will be used as an estimator. If $E \leq Tol$, y will be kept as the current step solution and then move to the next step with time step size δh. If $E > Tol$, recalculate the current step with time step size δh, where

$$\delta = 0.84\left(\frac{Tol}{E}\right)^{1/4}.$$

Problem 1

1. **The 4th RK method and RKF method for Problem 1.1**

   (a) **Results for Problem 1.1.** From the figure (Fig.1) we can see that the 4th RK method and RKF method are both convergent for Problem 1.1. The 4th RK method is convergent with 4 steps and RKF method with 2 steps and reached error 4.26×10^{-14}.

(b) **Figures** (Fig.1)

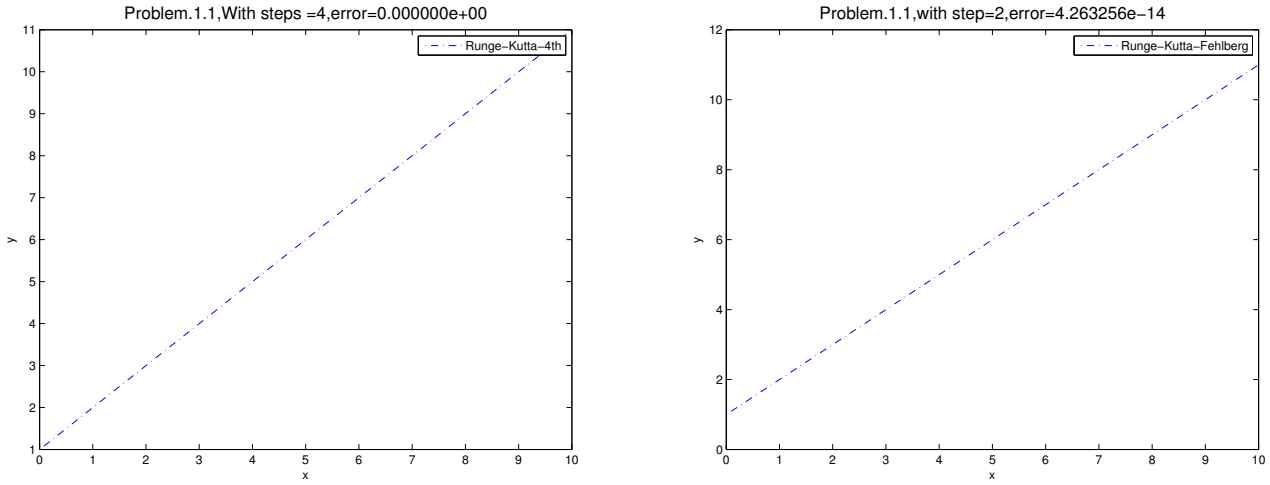

Figure 1: The 4th RK method and RKF method for Problem 1.1

2. **The 4th RK method and RKF method for Problem 1.2**

   (a) **Results for Problem 1.2.** From the figure (Fig.2) we can see that the 4th RK method and RKF method are both convergent for Problem 1.2. The 4th RK method is convergent with 404 steps and reached error 9.9×10^{-6}. RKF method with 29 steps and reached error 2.3×10^{-9}.

   (b) **Figures** (Fig.2)

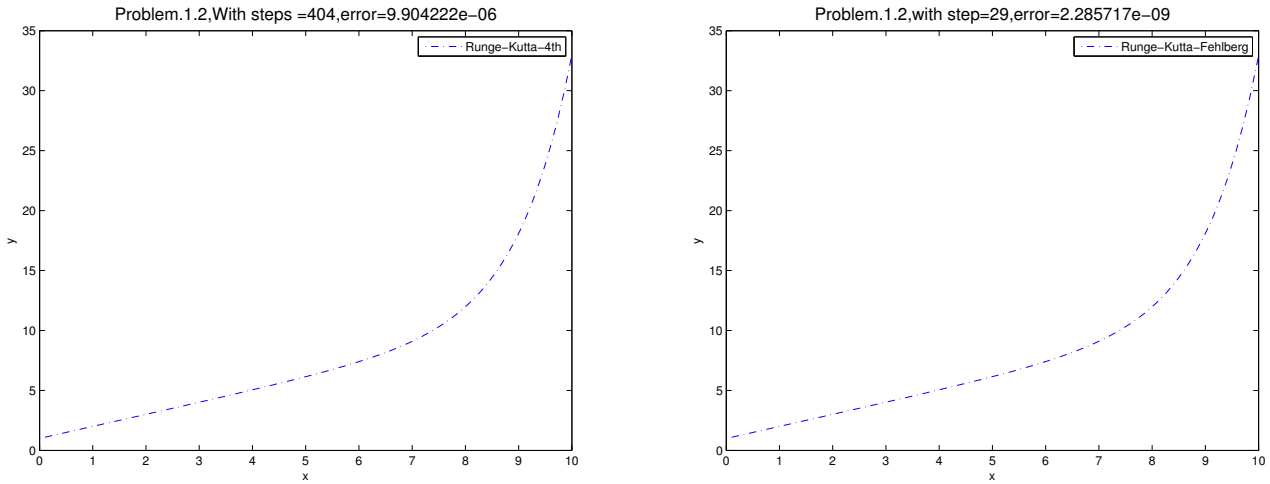

Figure 2: The 4th RK method and RKF method for Problem 1.2

Problem 2

1. **The 4th RK method and RKF method for Problem 2.1**

(a) **Results for Problem 2.1.** From the figure (Fig.3) we can see that the 4th RK method and RKF method are both convergent for Problem 2.1. The 4th RK method is convergent with 24 steps and reached error 7.1×10^{-6}. RKF method with 8 steps and reached error 9.4×10^{-10}.

(b) **Figures** (Fig.3)

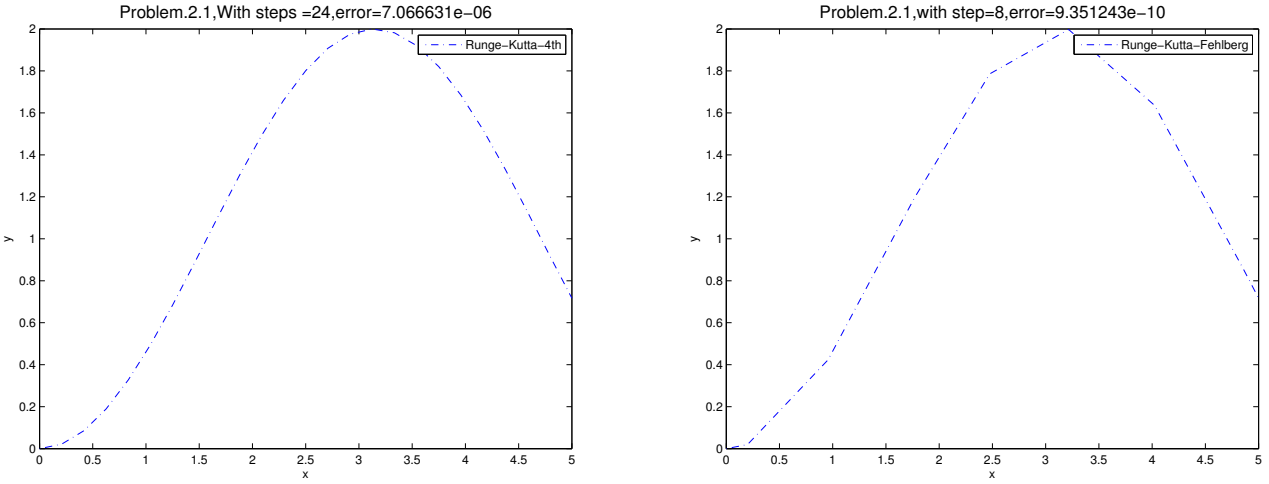

Figure 3: The 4th RK method and RKF method for Problem 2.1

2. **The 4th RK method and RKF method for Problem 2.2**

(a) **Results for Problem 2.2.** From the figure (Fig.4) we can see that the 4th RK method and RKF method are both divergent for Problem 2.2.

(b) **Figures** (Fig.4)

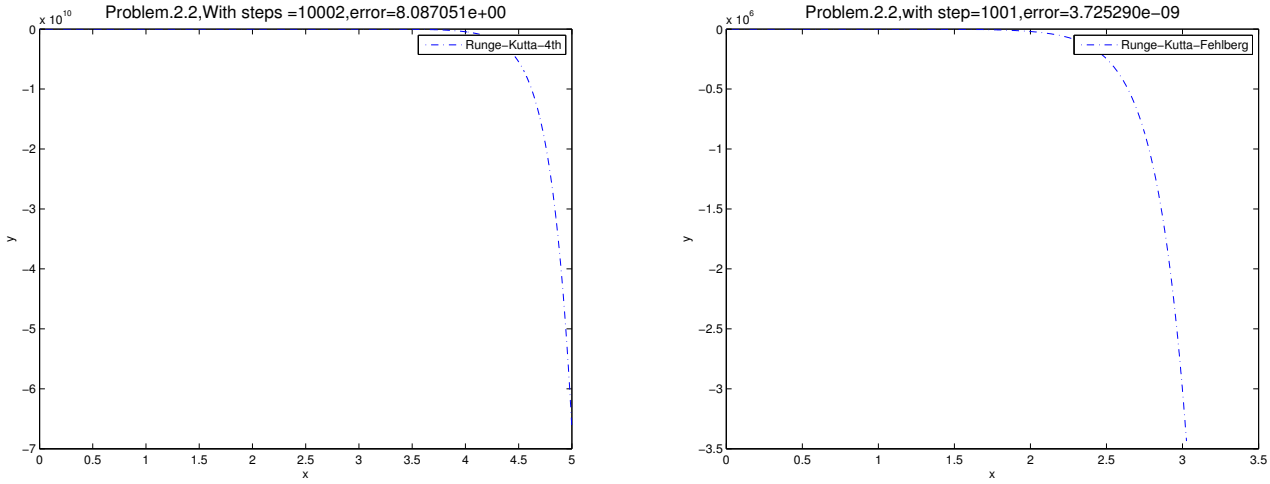

Figure 4: The 4th RK method and RKF method for Problem 2.2

3. **The 4th RK method and RKF method for Problem 2.3**

(a) **Results for Problem 2.3.** From the figure (Fig.5) we can see that the 4th RK method and RKF method are both convergent for Problem 2.3. The 4th RK method is convergent with 96 steps and reached error 9.98×10^{-6}. RKF method with 69 steps and reached error 1.3×10^{-11}.

(b) **Figures** (Fig.5)

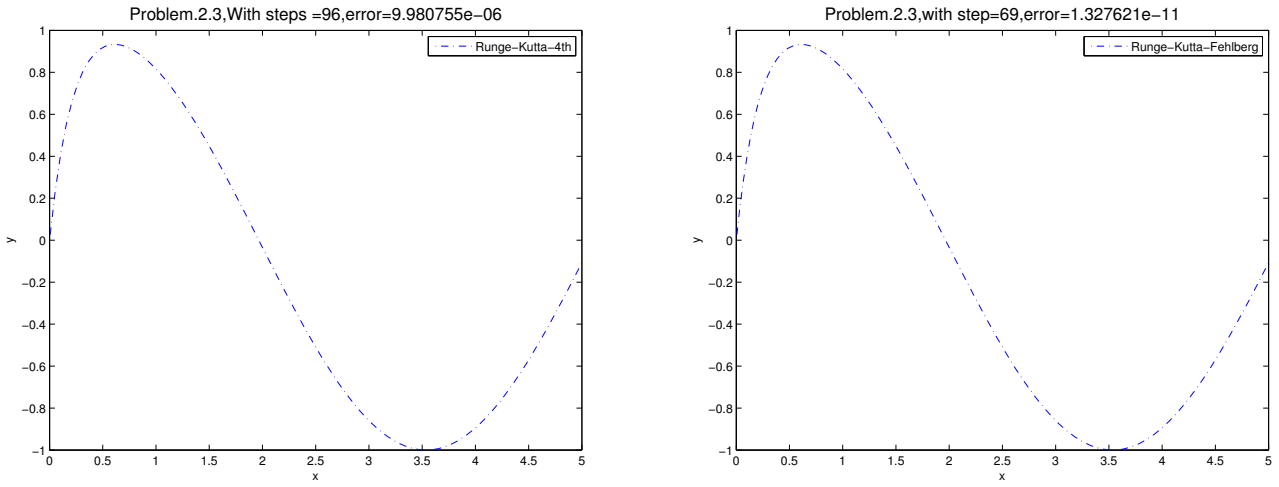

Figure 5: The 4th RK method and RKF method for Problem 2.3

(c) **The 4th RK method and RKF method for Problem 2.4**

i. **Results for Problem 2.4.** From the figure (Fig.6) we can see that the 4th RK method and RKF method are both divergent for Problem 2.4.

ii. **Figures** (Fig.6)

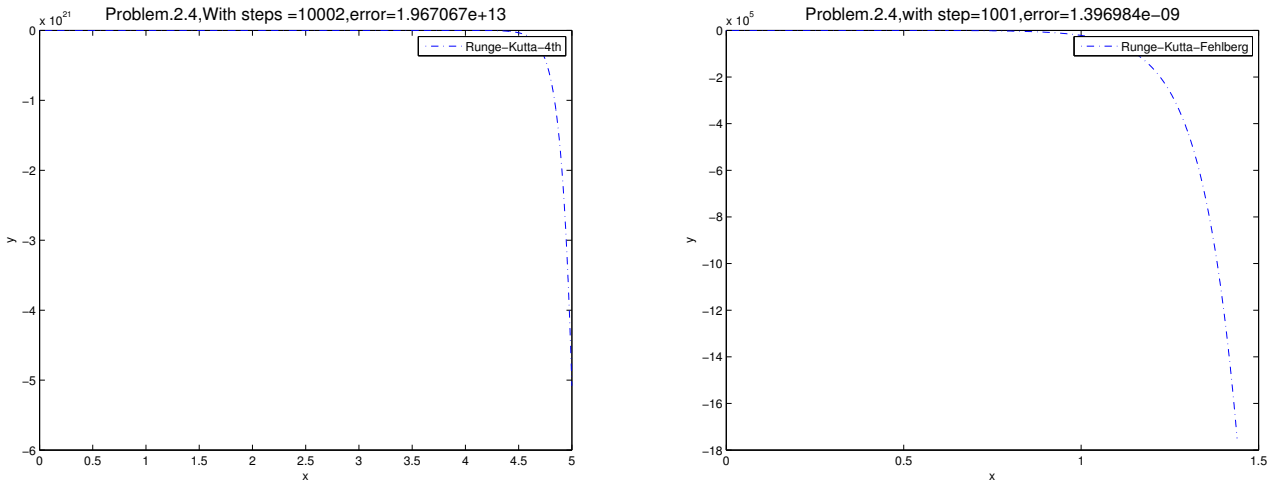

Figure 6: The 4th RK method and RKF method for Problem 2.4

(d) **The 4th RK method and RKF method for Problem 2.5**

i. **Results for Problem 2.5.** From the figure (Fig.7) we can see that the 4th RK method and RKF method are both convergent for Problem 2.5. The 4th RK method is convergent

with 88 steps and reached error 8.77×10^{-6}. RKF method with 114 steps and reached error 2.57×10^{-10}.

ii. **Figures** (Fig.7)

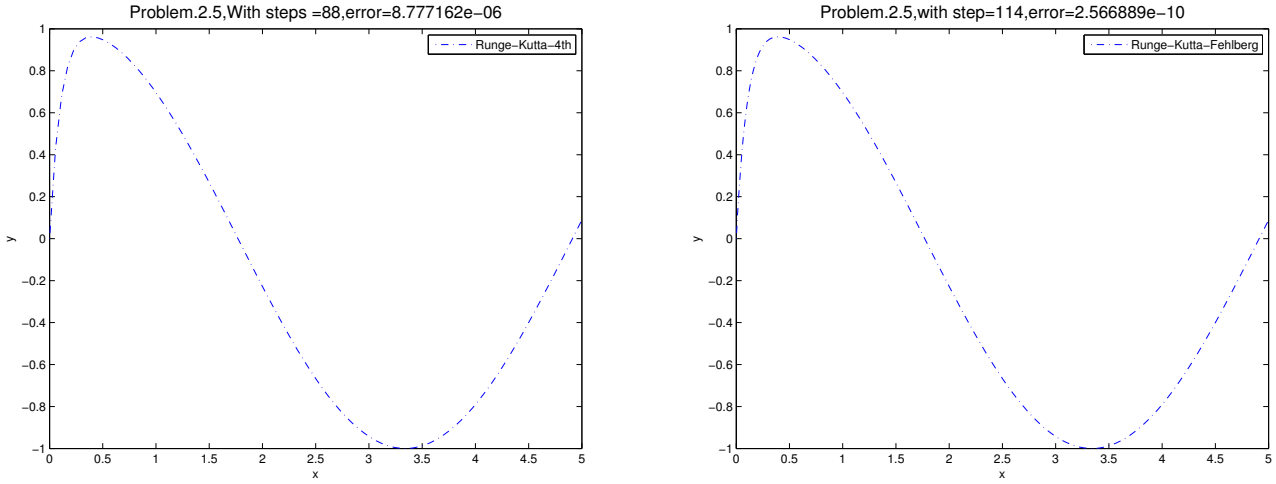

Figure 7: The 4th RK method and RKF method for Problem 2.5

Problem 3

1. **The 4th RK method and RKF method for Problem 3**

   (a) **Results for Problem 3.** From the figure (Fig.8) we can see that the 4th RK method and RKF method are both convergent for Problem 3. The 4th RK method is convergent with 4 steps and reached error 1.77×10^{-15}. RKF method with 2 steps and reached error 2×10^{-15}.

   (b) **Figures** (Fig.8)

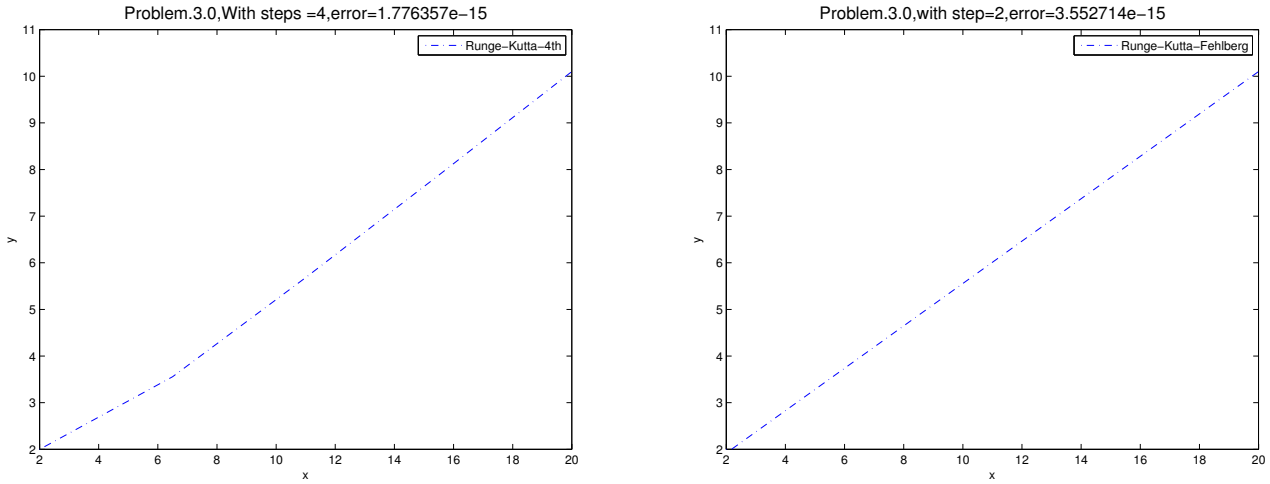

Figure 8: The 4th RK method and RKF method for Problem 3

Problem 4

1. **The 4th RK method and RKF method for Problem 4**

   (a) **Results for Problem 4.** From the figure (Fig.9) we can see that the 4th RK method and RKF method are both convergent for Problem 4. The 4th RK method is convergent with 438 steps and reached error 9.9×10^{-6}. RKF method with 134 steps and reached error 3.68×10^{-14}.

   (b) **Figures** (Fig.9)

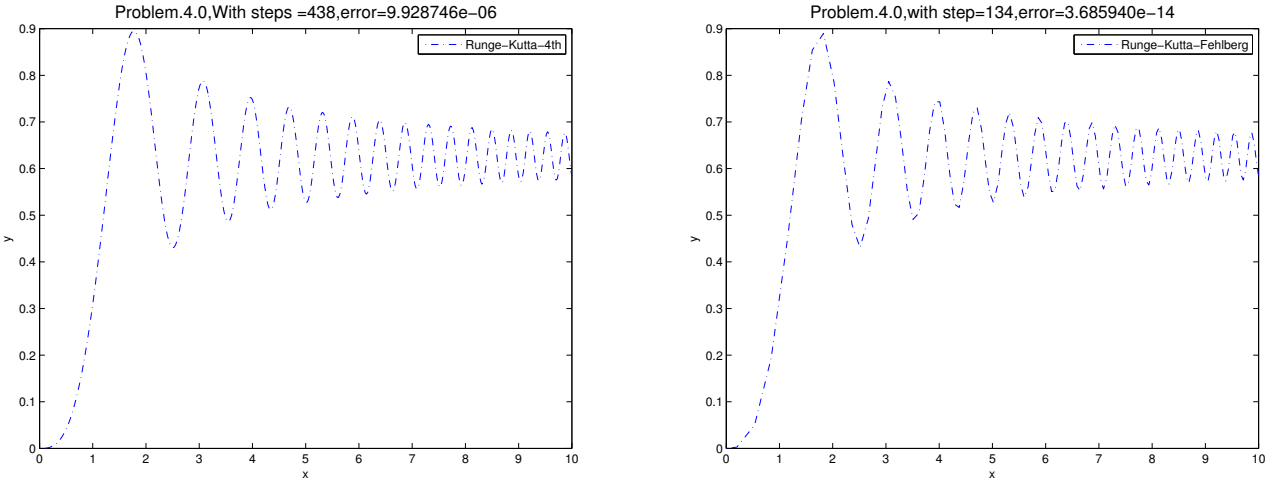

Figure 9: The 4th RK method and RKF method for Problem 4

Problem 5

1. **The 4th RK method and RKF method for Problem 5**

   (a) **Results for Problem 5.** Since, $x = 0$ is the singular point for the problems and $y_0 = \lim_{x \to 0^-} = 1$. So, the schemes do not work for the interval $[-2, 0]$. But schemes works for the interval $[-2, 0-\delta]$ and $\delta > 1 \times 10^{16}$. I changed the problem to the following

$$\begin{aligned} f'(x) &= \tfrac{\ln(1+x)}{x}, x \in [\delta, 2] \\ f(\delta) &= 0. \end{aligned}$$

   The (Fig.8) gives the result for the interval $[\delta, 2]$ and $\delta = 1 \times 10^{10}$.

   (b) **Figures** (Fig.10)

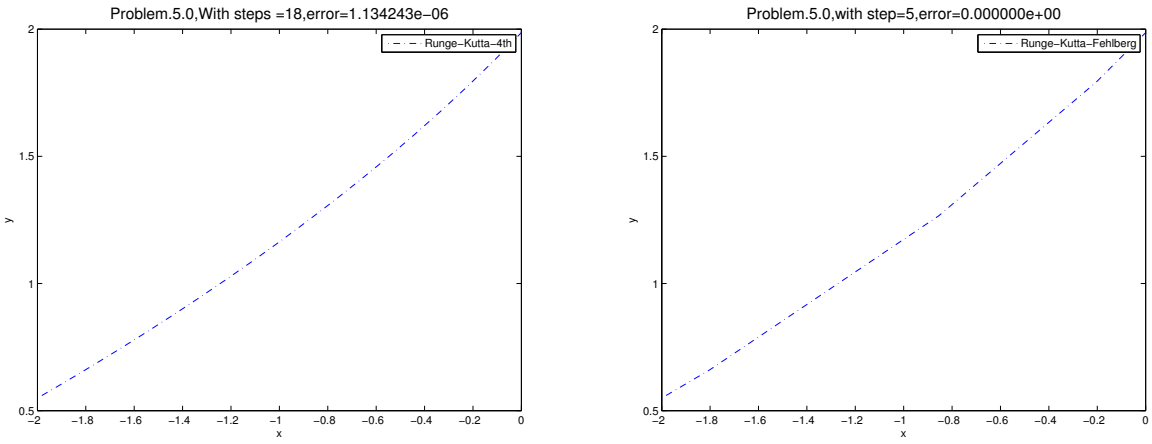

Figure 10: The 4th RK method and RKF method for Problem 5

Adaptive Runge-Kutta Methods MATLAB Code

1. **4-th oder Runge-Kutta Method MATLAB code**

Listing 1: 4-th oder Runge-Kutta Method

```
function [x,y,h]=Runge_Kutta_4(f,xinit,yinit,xfinal,n)
% Euler approximation for ODE initial value problem
% Runge-Kutta 4th order method
% author:Wenqiang Feng
% Email: fw253@mst.edu
% date:January 22, 2012
% Calculation of h from xinit, xfinal, and n
h=(xfinal-xinit)/n;
x=[xinit zeros(1,n)]; y=[yinit zeros(1,n)];

for i=1:n   %calculation loop
x(i+1)=x(i)+h;
k_1 = f(x(i),y(i));
k_2 = f(x(i)+0.5*h,y(i)+0.5*h*k_1);
k_3 = f((x(i)+0.5*h),(y(i)+0.5*h*k_2));
k_4 = f((x(i)+h),(y(i)+k_3*h));

y(i+1) = y(i) + (1/6)*(k_1+2*k_2+2*k_3+k_4)*h;  %main equation
end
```

2. **Main function for problems**

Listing 2: Main function for problem1-5 with 4-th oder Runge-Kutta Method

```
% Script file: main1.m
% The RHS of the differential equation is defined as
% a handle function
% author:Wenqiang Feng
% Email: wfeng1@utk.edu
% date: Mar 8, 2014
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%% common parameters
clc
clear all
n=1;
tol=1e-5;
choice=5;  % The choice of the problem number
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%% the parameters for each problems
switch choice
case 1.1
% problem 11
f=@(x,y) y-x; %The right hand term
xinit=0;
xfinal=10;
yinit=1;%+1e-3;  %The initial condition
case 1.2
% problem 12
```

```
f=@(x,y) y-x; %The right hand term
xinit=0;
xfinal=10;
yinit=1+1e-3;  %The initial condition
case 2.1
% problem 21
lambda=0;
f=@(x,y) lambda*y+sin(x)-lambda*cos(x); %The right hand term
xinit=0;
xfinal=5;
yinit=0;  %The initial condition
case 2.2
% problem 22
lambda=5;
f=@(x,y) lambda*y+sin(x)-lambda*cos(x); %The right hand term
xinit=0;
xfinal=5;
yinit=0;  %The initial condition
case 2.3
% problem 23
lambda=-5;
f=@(x,y) lambda*y+sin(x)-lambda*cos(x); %The right hand term
xinit=0;
xfinal=5;
yinit=0;  %The initial condition
case 2.4
% problem 24
lambda=10;
f=@(x,y) lambda*y+sin(x)-lambda*cos(x); %The right hand term
xinit=0;
xfinal=5;
yinit=0;  %The initial condition
case 2.5
% problem 25
lambda=-10;
f=@(x,y) lambda*y+sin(x)-lambda*cos(x); %The right hand term
xinit=0;
xfinal=5;
yinit=0;  %The initial condition
case 3
% problem 3
f=@(x,y) 1-y/x; %The right hand term
xinit=2;
xfinal=20;
yinit=2;  %The initial condition
case 4
% problem 4
f=@(x,y) sin(x^2); %The right hand term
xinit=0;
xfinal=10;
yinit=0;  %The initial condition

case 5
```

```
% problem 5
f=@(x,y) log(1+x)/x; %The right hand term
xinit=1e-10;
xfinal=2;
yinit=0;  %The initial condition
end


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%% computing the numberical solutions

y0=100*ones(1,n+1);
[x1,y1]=Runge_Kutta_4(f,xinit,yinit,xfinal,n);

% computing the initial error
%en=max(abs(y1-y0));
en=max(abs(y1(end)-y0(end)));
while (en>tol)
n=n+1;
[x1,y1]=Runge_Kutta_4(f,xinit,yinit,xfinal,n);
[x2,y2,h]=Runge_Kutta_4(f,xinit,yinit,xfinal,2*n);
% two method to computing the error
% temp=interp1(x1,y1,x2);
% en=max(abs(temp-y2));
en=max(abs(y1(end)-y2(end)));
if (n>5000)
disp('the partitions excess 1000')
break;
end
end
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%% Plot
figure
plot(x2,y2,'-.')
xlabel('x')
ylabel('y')
legend('Runge-Kutta-4th')
title(sprintf('Problem.%1.1f,With steps =%d,error=%1e',choice,2*n,en),...
'FontSize', 14)
```

3. **Adaptive Runge-Kutta-Fehlberg Method MATLAB code**

Listing 3: 4-th oder Runge-Kutta Method

```
function [time,u,i,E]=Runge_Kutta_Fehlberg(t,T,h,y,f,tol)
% author:Wenqiang Feng
% Email: wfeng1@utk.edu
% date: Mar 8, 2014
u0=y; % initial value
t0=t; % initial time
i=0;  % initial counter
while t<T
h = min(h, T-t);
k1 = h*f(t,y);
```

```
k2 = h*f(t+h/4, y+k1/4);
k3 = h*f(t+3*h/8, y+3*k1/32+9*k2/32);
k4 = h*f(t+12*h/13, y+1932*k1/2197-7200*k2/2197+7296*k3/2197);
k5 = h*f(t+h, y+439*k1/216-8*k2+3680*k3/513-845*k4/4104);
k6 = h*f(t+h/2, y-8*k1/27+2*k2-3544*k3/2565+1859*k4/4104-11*k5/40);
y1 = y + 16*k1/135+6656*k3/12825+28561*k4/56430-9*k5/50+2*k6/55;
y2 = y + 25*k1/216+1408*k3/2565+2197*k4/4104-k5/5;
E=abs(y1-y2);
R = E/h;
delta = 0.84*(tol/R)^(1/4);
if E<=tol
t = t+h;
y = y1;
i = i+1;
fprintf('Step %d: t = %6.4f, y = %18.15f\n', i, t, y);
u(i)=y;
time(i)=t;
h = delta*h;
else
h = delta*h;
end
if (i>1000)
disp('the partitions excess 1000')
break;
end
end
time=[t0,time];
u=[u0,u];
```

4. **Main function for problems**

Listing 4: Main function for problem1-5 with Adaptive Runge-Kutta-Fehlberg Method

```
%% main2
clc
clear all
%% common parameters
tol=1e-5;
h = 0.2;
choice=5;  % The choice of the problem number
%% the parameters for each problems
switch choice
case 1.1
% problem 11
f=@(x,y) y-x; %The right hand term
xinit=0;
xfinal=10;
yinit=1;%+1e-3;  %The initial condition
case 1.2
% problem 12
f=@(x,y) y-x; %The right hand term
xinit=0;
xfinal=10;
yinit=1+1e-3;  %The initial condition
```

```
case 2.1
% problem 21
lambda=0;
f=@(x,y) lambda*y+sin(x)-lambda*cos(x); %The right hand term
xinit=0;
xfinal=5;
yinit=0;  %The initial condition
case 2.2
% problem 22
lambda=5;
f=@(x,y) lambda*y+sin(x)-lambda*cos(x); %The right hand term
xinit=0;
xfinal=5;
yinit=0;  %The initial condition
case 2.3
% problem 23
lambda=-5;
f=@(x,y) lambda*y+sin(x)-lambda*cos(x); %The right hand term
xinit=0;
xfinal=5;
yinit=0;  %The initial condition
case 2.4
% problem 24
lambda=10;
f=@(x,y) lambda*y+sin(x)-lambda*cos(x); %The right hand term
xinit=0;
xfinal=5;
yinit=0;  %The initial condition
case 2.5
% problem 25
lambda=-10;
f=@(x,y) lambda*y+sin(x)-lambda*cos(x); %The right hand term
xinit=0;
xfinal=5;
yinit=0;  %The initial condition
case 3
% problem 3
f=@(x,y) 1-y/x; %The right hand term
xinit=2;
xfinal=20;
yinit=2;  %The initial condition
case 4
% problem 4
f=@(x,y) sin(x^2); %The right hand term
xinit=0;
xfinal=10;
yinit=0;  %The initial condition

case 5
% problem 5
f=@(x,y) log(1+x)/x; %The right hand term
xinit=1e-10;
xfinal=2;
```

```
yinit=0;  %The initial condition
end
% xinit = 0;
% xfinal=2;
% yinit = 0.5;
% f=@(t,y) y-t^2+1; %The right hand term

fprintf('Step %d: t = %6.4f, w = %18.15f\n', 0, xinit, yinit);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%% computing the numberical solutions
[time,u,step,error]=Runge_Kutta_Fehlberg(xinit,xfinal,h,yinit,f,tol);

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%% Plot
figure
plot(time,u,'-.')
xlabel('x')
ylabel('y')
legend('Runge-Kutta-Fehlberg')
title(sprintf('Problem.%1.1f,with step=%d,error=%1e',choice,step,error),...
'FontSize', 14)
```

G Project 2 MATH572

COMPUTATIONAL ASSIGNMENT #2

MATH 572

The purpose of this assignment is to explore techniques for the numerical solution of boundary value and initial boundary value problems and to introduce some ideas that we did not discuss in class but, nevertheless, are quite important. You should submit the solution of at least two (2) of the following problems. Submitting the solution to the third can be used for extra credit.

The Convection diffusion equation. Upwinding.

Let $\Omega = (0, 1)$ and consider the following two point boundary value problem:

$$-\epsilon u'' + u' = 0, \qquad u(0) = 1,\ u(1) = 0.$$

Here $\epsilon > 0$ is a constant. We are interested in what happens when $\epsilon \ll 1$.

- Find the exact solution to this problem. Is it monotone?
- Compute a finite difference approximation of this problem on a uniform grid of size $h = 1/N$ using centered differences: That is, set $U_0 = 1$, $U_N = 0$ and

$$\beta_i U_{i-1} + \alpha_i U_i + \gamma_i U_{i+1} = 0, \quad 0 < i < N, \tag{1}$$

  where $\alpha_i, \beta_i, \gamma_i$ are defined in terms of ϵ and h. Set $\epsilon \in \{1, 10^{-1}, 10^{-3}, 10^{-6}\}$ and compute the solution for different values of h. What do you observe for $h > \epsilon$? For $h \approx \epsilon$? For $h < \epsilon$?
- Show that

$$\hat{U}_i = 1, \qquad \check{U}_i = \left(\frac{\frac{2\epsilon}{h} + 1}{\frac{2\epsilon}{h} - 1}\right)^i, \qquad i = \overline{0, N}$$

  are two linearly independent solutions of the difference equation. Find the dicrete solution U of the problem in terms of $\hat{U}$ and $\check{U}$. Using this representation, determine the relation between ϵ and h that ensures that there are no oscillations in U. Does this coincide with your observations of the previous item?
  Hint: Consider the sign of $\frac{\frac{2\epsilon}{h}+1}{\frac{2\epsilon}{h}-1}$.
- Replace the centered difference approximation of the first derivative u' by the *up-wind difference* $u'(x_i) \approx h^{-1}(u(x_i) - u(x_{i-1}))$. Repeat the previous two items and draw conclusions.
- Show that, using an up-wind approximation the arising matrix satisfies a discrete maximum principle.

Date: Due April 24, 2014.

A posteriori error estimation

For this problem consider

$$-(a(x)u')' = f, \text{ in } (0,1), \quad u(0) = 0, \quad u(1) = 1. \tag{2}$$

Write a piece of code that, for a given a and f computes the finite element solution to this problem over a mesh $\mathcal{T}_h = \{I_j\}_{j=1}^N$, $I_j = [x_{j-1}, x_j]$ with $h_j = x_j - x_{j-1}$ not necessarily uniform.

- Set $a = 1$ and choose f so that $u = x^3$. Compute the finite element solution on a sequence of uniform meshes of size $h = 1/N$ and verify the estimate

$$\|u - U\|_{H^1(0,1)} \leq Ch = CN^{-1}. \tag{3}$$

- Set $a = 1$ and $f = -\frac{3}{4\sqrt{x}}$ and notice that $f \notin L^2(0,1)$. This problem, however, is still well posed. Show this. For this case repeat the previous item. What do you observe?
- Set $a(x) = 1$ if $0 \leq x < 1/\pi$ and $a(x) = 2$ otherwise. Choose $f \equiv 1$ and compute the exact solution. Repeat the first item. What do you observe? Recall that to compute the exact solution we must include the interface conditions: u and au' are continuous.

The last two items show that in the case when either the right hand side or the coefficient in the equation are not smooth, the solution does not satisfy $u'' \in L^2(0,1)$ and so the error estimate (3) cannot be obtained with uniform meshes. Notice, also, that in both cases the solution is smooth except perhaps at very few points, so that if we were able to handle these, problematic, points we should be able to recover (3). The purpose of *a posteriori* error estimates is exactly this.

Let us recall the weak formulation of (2). Define:

$$\mathcal{A}(v,w) = \int_0^1 av'w', \qquad \mathcal{L}(v) = \int_0^1 fv,$$

then we need to find u such that $u - 1 \in H_0^1(0,1)$ and

$$\mathcal{A}(u,v) = \mathcal{L}(v) \quad \forall v \in H_0^1(0,1).$$

If U is the finite element solution to (2) and $v \in H_0^1(0,1)$, we have

$$\mathcal{A}(u-U,v) = \mathcal{A}(u,v) - \mathcal{A}(U,v) = \mathcal{L}(v) - \mathcal{A}(U,v) = \int_0^1 fv - \int_0^1 aU'v' = \sum_{j=1}^N \int_{I_j} fv - aU'v'.$$

Let us now consider each integral separately. Integrating by parts we obtain

$$\int_{I_j} fv - aU'v' = \int_{I_j} fv + \int_{I_j} (aU')'v - aU'v\Big|_{x_{j-1}}^{x_j}$$

so that adding up we get

$$\mathcal{A}(u-U,v) = \sum_{j=1}^N \int_{I_j} (f + (aU')')\, v + \sum_{j=1}^{N-1} v(x_j)\mathbf{j}(a(x_j)U'(x_j)),$$

where

$$\mathbf{j}(w(x)) = w(x+0) - w(x-0)$$

is the so-called *jump*. Let us now set $v = w - I_h w$, where I_h is the Lagrange interpolation operator. In this case then $v(x_j) = 0$ (why?) and

$$\|v\|_{L^2(I_j)} = \|w - I_h w\|_{L^2(I_j)} \leq ch_j \|w'\|_{L^2(I_j)}.$$

Consequently,

$$\begin{aligned}\mathcal{A}(u-U, w-I_h w) &\le C \sum_{I_j\in\mathcal{T}_h} h_j \|f+(aU')'\|_{L^2(I_j)}\|w'\|_{L^2(I_j)} \\ &\le C\left(\sum_{I_j\in\mathcal{T}_h} h_j^2\|f+(aU')'\|^2_{L^2(I_j)}\right)^{1/2}\left(\sum_{I_j\in\mathcal{T}_h}\|w'\|_{L^2(I_j)}\right)^{1/2} \\ &= C\left(\sum_{I_j\in\mathcal{T}_h} h_j^2\|f+(aU')'\|^2_{L^2(I_j)}\right)^{1/2}\|w'\|_{L^2(0,1)}\end{aligned}$$

What is the use of all this? Define $\mathsf{r}_j = h_j\|f+(aU')'\|_{L^2(I_j)}$ then, using Galerkin orthogonality we obtain

$$\|u-U\|^2_{H^1(0,1)} \le \frac{1}{c_1}\mathcal{A}(u-U,u-U) = \frac{1}{c_1}\mathcal{A}(u-U,u-U-I_h(u-U)) \le C\left(\sum_{j=1}^N \mathsf{r}_j^2\right)^{1/2}\|u-U\|_{H^1(0,1)}.$$

In other words, we bounded the error in terms of **computable** and **local** quantities r_j. This allows us to devise an *adaptive* method:

- (Solve) Given $\mathcal{T}_h$ find U.
- (Estimate) Compute the r_j's.
- (Mark) Choose ℓ for which r_ℓ is maximal.
- (Refine) Construct a new mesh by *bisecting* I_ℓ and leaving all the other elements unchanged.

Implement this method and show that (3) is recovered.

You might also want to try choosing a set of minimal cardinality $\mathcal{M}$ so that $\sum_{j\in\mathcal{M}}\mathsf{r}_j^2 \ge \frac{1}{2}\sum_{j=1}^N \mathsf{r}_j^2$ and bisecting the cells I_j with $j\in\mathcal{M}$.

Numerical methods for the heat equation

Let $\Omega=(0,1)$ and $T=1$. Consider the heat equation

$$u_t - u'' = f \text{ in } \Omega, \qquad u|_{\partial\Omega}=0, \qquad u|_{t=0}=u_0.$$

Choose f and u_0 so that the exact solution reads:

$$u(x,t)=\sin(3\pi x)e^{-2t},$$

Implement a finite difference discretization of this problem in sapce and, in time, the θ-method:

$$\frac{U_i^{k+1}-U_i^k}{\tau} - \theta\Delta_h U_i^k - (1-\theta)\Delta_h U_i^{k+1} = f_i^{k+1}.$$

In doing so you obtained:

- The explicit Euler method, $\theta=1$.
- The implicit Euler method, $\theta=0$.
- The Crank-Nicolson method, $\theta=\frac{1}{2}$.

For each one of them compute the discrete solution U at $T=1$ and measure the L^2, H^1 and L^∞ norms of the error. You should do this on a series of meshes and verify the theoretical error estimates. The time step must be chosen as:

- $\tau=\sqrt{h}$.
- $\tau=h$.
- $\tau=h^2$.

What can you conclude?

MATH 572: Computational Assignment #2

Due on Thurday, April 24, 2014

TTH 12:40pm

Wenqiang Feng

Wenqiang Feng MATH 572 (TTH 12:40pm): Computational Assignment #2

Contents

The Convection Diffusion Equation

Problem 1

1. **The exact solution**

   From the problem, we know that the characteristic function is

   $$-\epsilon\lambda^2 + \lambda = 0.$$

   So, $\lambda = 0, \frac{1}{\epsilon}$. Therefore, the general solution is

   $$u = c_1 e^{0x} + c_2 e^{\frac{1}{\epsilon}x} = c_1 + c_2 e^{\frac{1}{\epsilon}x}.$$

   By using the boundary conditions, we get the solution is

   $$u(x) = 1 - \frac{1}{1-e^{\frac{1}{\epsilon}}} + \frac{1}{1-e^{\frac{1}{\epsilon}}} e^{\frac{1}{\epsilon}x}.$$

   And $u(x)$ is monotone.

2. **Central Finite difference scheme**

   I consider the following partition for finite difference method:

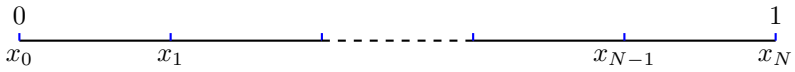

Figure 1: One dimension's uniform partition for finite difference method

   Then, the central difference scheme is as following:

$$-\epsilon\frac{U_{i-1} - 2U_i + U_{i+1}}{h^2} + \frac{U_{i+1} - U_{i-1}}{2h} = 0, \quad i = 1, 2, \cdots, N-1. \tag{1}$$

$$U_0 = 1, U_N = 0. \tag{2}$$

   So

   (a) when $i = 1$, we get

$$-\epsilon\frac{U_0 - 2U_1 + U_2}{h^2} + \frac{U_2 - U_0}{2h} = 0,$$

   i.e.

$$-\left(\frac{\epsilon}{h^2} + \frac{1}{2h}\right)U_0 + \frac{2\epsilon}{h^2}U_1 + \left(\frac{1}{2h} - \frac{\epsilon}{h^2}\right)U_2 = 0.$$

   Since, $U_0 = 1$, so we get

$$\frac{2\epsilon}{h^2}U_1 + \left(\frac{1}{2h} - \frac{\epsilon}{h^2}\right)U_2 = \left(\frac{\epsilon}{h^2} + \frac{1}{2h}\right). \tag{3}$$

   (b) when $i = 2, \cdots, N-2$, we get

$$-\epsilon\frac{U_{i-1} - 2U_i + U_{i+1}}{h^2} + \frac{U_{i+1} - U_{i-1}}{2h} = 0.$$

   i.e.

$$-\left(\frac{\epsilon}{h^2} + \frac{1}{2h}\right)U_{i-1} + \frac{2\epsilon}{h^2}U_i + \left(\frac{1}{2h} - \frac{\epsilon}{h^2}\right)U_{i+1} = 0. \tag{4}$$

3. when $i = N-1$

$$-\epsilon\frac{U_{N-2}-2U_{N-1}+U_N}{h^2}+\frac{U_N-U_{N-2}}{2h}=0,$$

i.e.

$$-\left(\frac{\epsilon}{h^2}+\frac{1}{2h}\right)U_{N-2}+\frac{2\epsilon}{h^2}U_{N-1}+\left(\frac{1}{2h}-\frac{\epsilon}{h^2}\right)U_N=0.$$

Since $U_N=0$, then,

$$-\left(\frac{\epsilon}{h^2}+\frac{1}{2h}\right)U_{N-2}+\frac{2\epsilon}{h^2}U_{N-1}=0. \tag{5}$$

From (3)-(5), we get the algebraic system is

$$AU=F,$$

where

$$A=\begin{pmatrix} \frac{2\epsilon}{h^2} & \frac{1}{2h}-\frac{\epsilon}{h^2} & & & \\ -\left(\frac{\epsilon}{h^2}+\frac{1}{2h}\right) & \frac{2\epsilon}{h^2} & \frac{1}{2h}-\frac{\epsilon}{h^2} & & \\ & \ddots & \ddots & \ddots & \\ & & -\left(\frac{\epsilon}{h^2}+\frac{1}{2h}\right) & \frac{2\epsilon}{h^2} & \frac{1}{2h}-\frac{\epsilon}{h^2} \\ & & & -\left(\frac{\epsilon}{h^2}+\frac{1}{2h}\right) & \frac{2\epsilon}{h^2} \end{pmatrix},$$

$$U=\begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_{N-2} \\ U_{N-1} \end{pmatrix}, F=\begin{pmatrix} \frac{\epsilon}{h^2}+\frac{1}{2h} \\ \vdots \\ 0 \\ \vdots \end{pmatrix}.$$

4. **Numerical Results of Central Difference Method**

h	N_{nodes}	$\|u-u_h\|_{l^\infty,\epsilon=1}$	$\|u-u_h\|_{l^\infty,\epsilon=10^{-1}}$	$\|u-u_h\|_{l^\infty,\epsilon=10^{-3}}$	$\|u-u_h\|_{l^\infty,\epsilon=10^{-6}}$
0.5	3	2.540669×10^{-3}	7.566929×10^{-1}	1.245000×10^{2}	∞
0.25	5	6.175919×10^{-4}	1.933238×10^{-1}	3.050403×10^{1}	∞
0.125	9	1.563835×10^{-4}	5.570936×10^{-2}	7.449173×10^{0}	∞
0.0625	17	3.928711×10^{-5}	1.211929×10^{-2}	1.692902×10^{0}	∞
0.03125	33	9.827515×10^{-6}	3.018484×10^{-3}	2.653958×10^{-1}	∞
0.015625	65	2.457936×10^{-6}	7.484336×10^{-4}	7.515267×10^{-3}	∞
0.007812	129	6.144675×10^{-7}	1.870750×10^{-4}	2.281210×10^{-9}	∞
0.003906	257	1.536257×10^{-7}	4.674564×10^{-5}	6.661338×10^{-16}	∞

Table 1: l^∞ norms for the Central Difference Method with $\epsilon=\{1,10^{-1},10^{-3},10^{-6}\}$

From Table.1, we get that

(a) when $h<\epsilon$ the scheme is convergent with optimal convergence order (Figure.2), i.e.

$$\|u-u_h\|_{l^\infty}\approx 0.01h^{1.9992},$$

(b) when $h \approx \epsilon$ the scheme is convergent with optimal convergence order (Figure.2), i.e.

$$\|u - u_h\|_{l^\infty} \approx 3.201 h^{2.0072},$$

(c) when $h > \epsilon$ the scheme is not stable and the solution has oscillation.

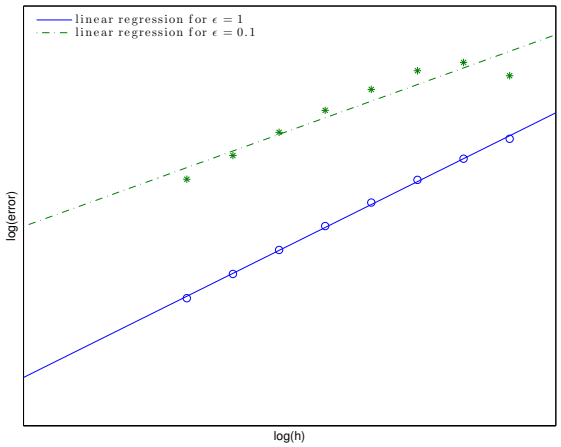

Figure 2: linear regression for l^∞ norm with $\epsilon = 1$ and $\epsilon = 0.1$

5. **Linearly Independent Solutions** $\hat{U}_i$ $\check{U}_i$

(a) Linearity
It is easy to check that

$$C_1 \hat{U}_i + C_2 \check{U}_i = 0,$$

only when $C_1 = C_2 = 0$.

(b) Solutions to (4)
Checking for $\hat{U}_i = 1$

$$-\epsilon \frac{1 - 2 * 1 + 1}{h^2} + \frac{1 - 1}{2h} = 0$$

Checking for $\check{U}_i = \left(\frac{\frac{2\epsilon}{h}+1}{\frac{2\epsilon}{h}-1}\right)^i$

$$\begin{aligned}
& - \epsilon \frac{\left(\frac{\frac{2\epsilon}{h}+1}{\frac{2\epsilon}{h}-1}\right)^{i-1} - 2\left(\frac{\frac{2\epsilon}{h}+1}{\frac{2\epsilon}{h}-1}\right)^{i} + \left(\frac{\frac{2\epsilon}{h}+1}{\frac{2\epsilon}{h}-1}\right)^{i+1}}{h^2} + \frac{\left(\frac{\frac{2\epsilon}{h}+1}{\frac{2\epsilon}{h}-1}\right)^{i+1} - \left(\frac{\frac{2\epsilon}{h}+1}{\frac{2\epsilon}{h}-1}\right)^{i-1}}{2h} \\
= & -\left(\frac{\epsilon}{h^2} + \frac{1}{2h}\right)\left(\frac{\frac{2\epsilon}{h}+1}{\frac{2\epsilon}{h}-1}\right)^{i-1} + \frac{2\epsilon}{h^2}\left(\frac{\frac{2\epsilon}{h}+1}{\frac{2\epsilon}{h}-1}\right)^{i} + \left(\frac{1}{2h} - \frac{\epsilon}{h^2}\right)\left(\frac{\frac{2\epsilon}{h}+1}{\frac{2\epsilon}{h}-1}\right)^{i+1} \\
= & -\frac{2\epsilon + h}{2h^2}\left(\frac{2\epsilon + h}{2\epsilon - h}\right)^{i-1} + \frac{2\epsilon}{h^2}\left(\frac{2\epsilon + h}{2\epsilon - h}\right)^{i} + \frac{h - 2\epsilon}{2h^2}\left(\frac{2\epsilon + h}{2\epsilon - h}\right)^{i+1} \\
= & -\frac{2\epsilon - h}{2h^2}\left(\frac{2\epsilon + h}{2\epsilon - h}\right)^{i} + \frac{2\epsilon}{h^2}\left(\frac{2\epsilon + h}{2\epsilon - h}\right)^{i} - \frac{2\epsilon + h}{2h^2}\left(\frac{2\epsilon + h}{2\epsilon - h}\right)^{i} \\
= & -\frac{2\epsilon}{h^2}\left(\frac{2\epsilon + h}{2\epsilon - h}\right)^{i} + \frac{2\epsilon}{h^2}\left(\frac{2\epsilon + h}{2\epsilon - h}\right)^{i} = 0
\end{aligned}$$

(c) The representation of $\hat{U}$ and $\check{U}$

Since $\hat{U}$ and $\check{U}$ are the solution of 1, so the linear combination is also solution to 1, i.e.

$$u = c_1\hat{U} + c_2\check{U}$$

is also solution to 1. We also need this solution to satisfy the boundary conditions, so

$$\begin{cases} u = c_1 + c_2\left(\frac{\frac{2\epsilon}{h}+1}{\frac{2\epsilon}{h}-1}\right) = 1 \\ u = c_1 + c_2\left(\frac{\frac{2\epsilon}{h}+1}{\frac{2\epsilon}{h}-1}\right)^N = 0. \end{cases}$$

so

$$c_1 = -\frac{(2\epsilon+h)^N}{(2\epsilon+h)(2\epsilon-h)^{N-1}-(2\epsilon+h)^N}, c_2 = \frac{(2\epsilon-h)^N}{(2\epsilon+h)(2\epsilon-h)^{N-1}-(2\epsilon+h)^N}.$$

6. **Up-wind Finite difference scheme**

By using the same partition as central difference, then the up-wind difference scheme is as following:

$$\begin{aligned} -\epsilon\frac{U_{i-1}-2U_i+U_{i+1}}{h^2} + \frac{U_i - U_{i-1}}{h} &= 0, \quad i = 1, 2, \cdots, N-1. \\ U_0 = 1, U_N &= 0. \end{aligned}$$

So

(a) when $i = 1$, we get

$$-\epsilon\frac{U_0-2U_1+U_2}{h^2} + \frac{U_1-U_0}{h} = 0,$$

i.e.

$$-\left(\frac{\epsilon}{h^2}+\frac{1}{h}\right)U_0 + \left(\frac{2\epsilon}{h^2}+\frac{1}{h}\right)U_1 - \frac{\epsilon}{h^2}U_2 = 0.$$

Since, $U_0 = 1$, so we get

$$\left(\frac{2\epsilon}{h^2}+\frac{1}{h}\right)U_1 - \frac{\epsilon}{h^2}U_2 = \left(\frac{\epsilon}{h^2}+\frac{1}{h}\right). \tag{6}$$

(b) when $i = 2, \cdots, N-2$, we get

$$-\epsilon\frac{U_{i-1}-2U_i+U_{i+1}}{h^2} + \frac{U_i-U_{i-1}}{h} = 0.$$

i.e.

$$-\left(\frac{\epsilon}{h^2}+\frac{1}{h}\right)U_{i-1} + \left(\frac{2\epsilon}{h^2}+\frac{1}{h}\right)U_i - \frac{\epsilon}{h^2}U_{i+1} = 0. \tag{7}$$

(c) when $i = N-1$

$$-\epsilon\frac{U_{N-2}-2U_{N-1}+U_N}{h^2} + \frac{U_{N-1}-U_{N-2}}{h} = 0,$$

i.e.

$$-\left(\frac{\epsilon}{h^2}+\frac{1}{h}\right)U_{N-2} + \left(\frac{2\epsilon}{h^2}+\frac{1}{h}\right)U_{N-1} - \frac{\epsilon}{h^2}U_N = 0.$$

Since $U_N = 0$, then,

$$-\left(\frac{\epsilon}{h^2}+\frac{1}{h}\right)U_{N-2}+\left(\frac{2\epsilon}{h^2}+\frac{1}{h}\right)U_{N-1}=0. \tag{8}$$

From (6)-(8), we get the algebraic system is

$$AU = F,$$

where

$$A=\begin{pmatrix} \left(\frac{2\epsilon}{h^2}+\frac{1}{h}\right) & -\frac{\epsilon}{h^2} & & & \\ -\left(\frac{\epsilon}{h^2}+\frac{1}{h}\right) & \left(\frac{2\epsilon}{h^2}+\frac{1}{h}\right) & -\frac{\epsilon}{h^2} & & \\ & \ddots & \ddots & \ddots & \\ & & -\left(\frac{\epsilon}{h^2}+\frac{1}{h}\right) & \left(\frac{2\epsilon}{h^2}+\frac{1}{h}\right) & -\frac{\epsilon}{h^2} \\ & & & -\left(\frac{\epsilon}{h^2}+\frac{1}{h}\right) & \left(\frac{2\epsilon}{h^2}+\frac{1}{h}\right) \end{pmatrix},$$

$$U=\begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_{N-2} \\ U_{N-1} \end{pmatrix}, F=\begin{pmatrix} \frac{\epsilon}{h^2}+\frac{1}{h} \\ \vdots \\ 0 \\ \vdots \end{pmatrix}.$$

7. **Numerical Results of Up-wind Difference Scheme**

h	N_{nodes}	$\|u-u_h\|_{l^\infty,\epsilon=1}$	$\|u-u_h\|_{l^\infty,\epsilon=10^{-1}}$	$\|u-u_h\|_{l^\infty,\epsilon=10^{-3}}$	$\|u-u_h\|_{l^\infty,\epsilon=10^{-6}}$
0.5	3	2.245933×10^{-2}	1.361643×10^{-1}	1.992032×10^{-3}	∞
0.25	5	1.270323×10^{-2}	1.988791×10^{-1}	1.587251×10^{-5}	∞
0.125	9	6.925118×10^{-3}	1.571250×10^{-1}	4.999060×10^{-7}	∞
0.0625	17	3.623644×10^{-3}	9.196290×10^{-2}	9.685710×10^{-10}	∞
0.03125	33	1.849028×10^{-3}	5.061410×10^{-2}	1.110223×10^{-15}	∞
0.015625	65	9.343457×10^{-4}	2.695432×10^{-2}	2.220446×10^{-16}	∞
0.007812	129	4.695265×10^{-4}	1.391029×10^{-2}	1.554312×10^{-15}	∞
0.003906	257	2.353710×10^{-4}	7.064951×10^{-3}	8.881784×10^{-16}	∞

Table 2: l^∞ norms for the Up-wind Difference Method with $\epsilon = \{1, 10^{-1}, 10^{-3}, 10^{-6}\}$

From the Table.2 we get that

(a) when $h < \epsilon$ the scheme is convergent with optimal convergence order (Figure.2), i.e.

$$\|u-u_h\|_{l^\infty} \approx 0.0471 h^{0.946},$$

(b) when $h \approx \epsilon$ the scheme is convergent, but the convergence order is not optimal (Figure.2), i.e.

$$\|u-u_h\|_{l^\infty} \approx 0.4398 h^{0.6852},$$

(c) when $h > \epsilon$ the scheme is convergent, and the solution has no oscillation.

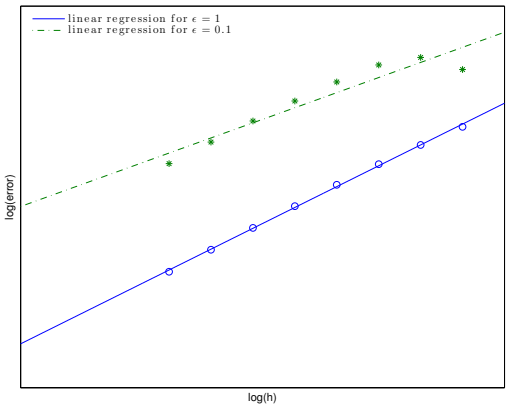

Figure 3: linear regression for l^∞ norm with $\epsilon = 1$ and $\epsilon = 0.1$

8. **Maximum Principle of Up-wind Difference Scheme**

   Lemma 0.1 *Let $A = tridiag\{a_i, b_i, c_i\}_{i=1}^n \in \mathbb{R}^{n\times n}$ be a tridiagional matrix with the properties that*

   $$b_i > 0, \quad a_i, c_i \leq 0, \quad a_i + b_i + c_i = 0.$$

   Then the following maximum principle holds: If $u \in \mathbb{R}^n$ is such that $(Au)_{i=2,\cdots,n-1} \leq 0$, then $u_i \leq \max\{u_1, u_n\}$.

   From the Up-wind Difference scheme, we get that $a_1 = 0$, $a_i = -\left(\frac{\epsilon}{h^2} + \frac{1}{h}\right), i = 2, \cdots, n$, $b_i = \left(\frac{2\epsilon}{h^2} + \frac{1}{h}\right), i = 1, \cdots, n$ and $c_i = -\frac{\epsilon}{h^2}, i = 1, \cdots, n-1$, moreover $(Au)_{i=2,\cdots,n-1} = 0$. Therefore,

   $$b_i > 0, \quad a_i, c_i \leq 0, \quad a_i + b_i + c_i = 0.$$

   Since $(Au)_{i=2,\cdots,n-1} = 0$, so the corresponding matrix arising from the up-wind scheme satisfies the discrete maximum principle(Lemma 0.1).

A Posterior Error Estimation

Problem 2

1. **Partition**

   I consider the following partition for finite element method:

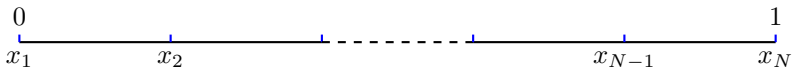

Figure 4: One dimension's uniform partition for finite element method

2. **Basis Function**

I will use the linear basis function, i.e. for each element $I = [x_i, x_{i+1}]$

$$\phi_I(x) = \begin{cases} \phi_1(x) = \frac{x_{i+1}-x}{x_{i+1}-x_i} \\ \phi_2(x) = \frac{x-x_i}{x_{i+1}-x_i}. \end{cases}$$

3. **Weak Formula**

   Multiplying the testing function $v \in H_0^1$ to both side of the problem, then integrating by part we get the following weak formula

$$\int_0^1 a(x)u'v'dx = \int_0^1 fvdx.$$

4. **Approximate Problem**

   The approximate problem is to find $u_h \in H^1$, s.t

$$a(u_h, v_h) = f(v_h) \forall v \in H_0^1,$$

   where

$$a(u_h, v_h) = \int_0^1 a(x)u_h'v_h'dx \quad \text{and} \quad f(v_h) = \int_0^1 fv_hdx.$$

5. **Numerical Results of Finite Element Method for Poisson Equation**

   (a) Problem: a(x)=1, $u_e = x^3$ and $f = -6x$.

h	N_{nodes}	$\|u-u_h\|_{L^2}$	$\|u-u_h\|_{H^1}$
1/4	5	1.791646×10^{-2}	2.480392×10^{-1}
1/8	9	4.502711×10^{-3}	1.247556×10^{-1}
1/16	17	1.127148×10^{-3}	6.246947×10^{-2}
1/32	33	2.818787×10^{-4}	3.124619×10^{-2}
1/64	65	7.047542×10^{-5}	1.562452×10^{-2}
1/128	128	1.761921×10^{-5}	7.812440×10^{-3}
1/256	257	4.404826×10^{-6}	3.906243×10^{-3}

Table 3: L^2 and H^1 Errors of Finite Element Method for Poisson Equation .

Using linear regression (Figure.5), we can also see that the errors in Table.4 obey

$$\|u-u_h\|_{L^2} \approx 0.2870h^{1.9987},$$
$$\|u-u_h\|_{H^1} \approx 0.9935h^{0.9986}.$$

These linear regressions indicate that the finite element method for this problem can converge in the optimal rates, which are second order in L^2 norm and first order in H^1 norm.

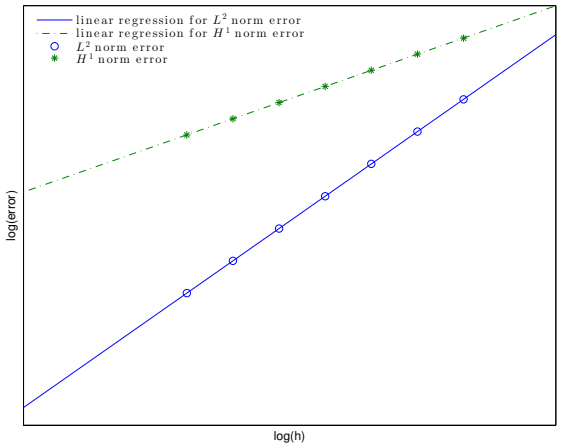

Figure 5: linear regression for L^2 and H^1 norm errors

(b) Problem: a(x)=1, $u_e = x^{\frac{3}{2}}$ and $f = -\frac{3}{4\sqrt{x}}$.

h	N_{nodes}	$\|u-u_h\|_{L^2}$	$\|u-u_h\|_{H^1}$
1/4	5	7.625472×10^{-3}	1.022294×10^{-1}
1/8	9	2.029299×10^{-3}	5.585353×10^{-2}
1/16	17	5.324774×10^{-4}	3.011300×10^{-2}
1/32	33	1.378846×10^{-4}	1.607571×10^{-2}
1/64	65	3.523180×10^{-5}	8.517032×10^{-3}
1/128	128	8.876332×10^{-6}	4.485323×10^{-3}
1/256	257	2.203920×10^{-6}	2.350599×10^{-3}

Table 4: L^2 and H^1 Errors of Finite Element Method for Poisson Equation .

Using linear regression (Figure.6), we can also see that the errors in Table.4 obey

$$\|u-u_h\|_{L^2} \approx 0.1193 h^{1.9593},$$
$$\|u-u_h\|_{H^1} \approx 0.3682 h^{0.9081}.$$

These linear regressions indicate that the finite element method for this problem can converge, but not in the optimal rates.

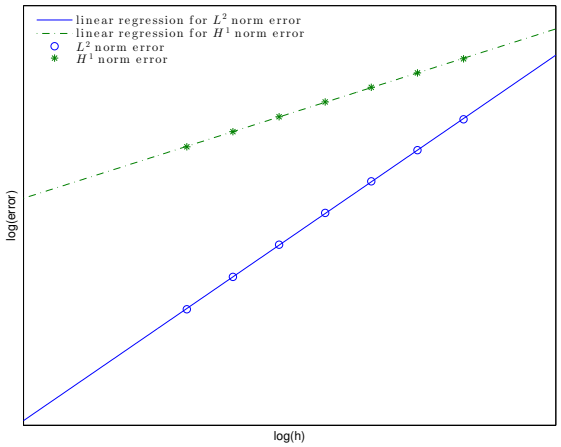

Figure 6: linear regression for L^2 and H^1 norm errors

(c) Problem: f=1,

$$a(x) = \begin{cases} 1, & 0 \le x < \frac{1}{\pi} \\ 2, & \frac{1}{\pi} \le x \le 1. \end{cases}$$

So, the exact solution should be

$$u_e = \begin{cases} -\frac{1}{2}x^2 + \frac{5\pi^2+1}{2\pi(\pi+1)}x, & 0 \le x < \frac{1}{\pi} \\ -\frac{1}{4}x^2 + \frac{5\pi^2+1}{4\pi(\pi+1)}x + \frac{5\pi-1}{4\pi(\pi+1)}, & \frac{1}{\pi} \le x \le 1. \end{cases}$$

We can not use the uniform mesh to compute this problem. Since if we can use the uniform mesh, then $\frac{1}{\pi}$ should be the node point, that is to say

$$nh = n\frac{1}{N_{elem}} = \frac{1}{\pi},$$

i.e.

$$n\pi = N_{elem}, n, N_{elem} \in \mathbb{Z}.$$

This is not possible, so we can not generate such mesh.

6. **Adaptive Finite Element Method for Possion Equation**

I will follow the standard local mesh refinement loops :

$$\text{SOLVE} \to \text{ESTIMATE} \to \text{MARK} \to \text{REFINE}.$$

(a) Problem: a(x)=1, $u_e = x^{\frac{3}{2}}$ and $f = -\frac{3}{4\sqrt{x}}$.

Iter	N_{elem}	$\|u-u_h\|_{L^\infty}$	$\|u-u_h\|_{L^2}$	$\|u-u_h\|_{H^1}$
1	32	2.797720×10^{-5}	1.378846×10^{-4}	1.607571×10^{-2}
2	47	1.022508×10^{-5}	6.093669×10^{-5}	9.927148×10^{-3}
3	75	3.674022×10^{-6}	2.038303×10^{-5}	5.935496×10^{-3}
4	102	1.313414×10^{-6}	1.400631×10^{-5}	4.239849×10^{-3}
5	145	4.663453×10^{-7}	6.119733×10^{-6}	2.933869×10^{-3}
6	171	1.654010×10^{-7}	4.589394×10^{-6}	2.432512×10^{-3}
7	192	5.970786×10^{-8}	4.010660×10^{-6}	2.185324×10^{-3}
8	208	5.956431×10^{-8}	3.587483×10^{-6}	2.034418×10^{-3}
9	219	5.957050×10^{-8}	3.297922×10^{-6}	1.942123×10^{-3}
10	229	5.976916×10^{-8}	3.076573×10^{-6}	1.864147×10^{-3}

Table 5: L^2 and H^1 Errors of Finite Element Method for Poisson Equation .

Using linear regression, we can also see that the errors (Figure.7) in Table.5 obey

$$\|u-u_h\|_{H^1} \approx 0.6454 N_{elem}^{-1.0798}.$$

These linear regressions indicate that the adaptive finite element method for this problem can converge in the optimal rates, which is first order in H^1 norm.

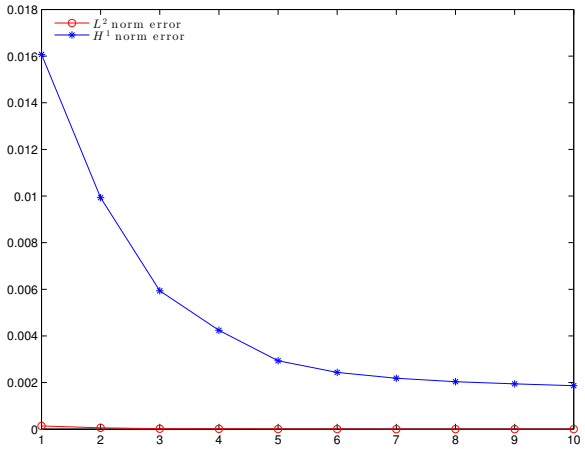

Figure 7: L^2 and H^1 norm errors for each iteration

(b) Problem: f=1,

$$a(x) = \begin{cases} 1, & 0 \le x < \frac{1}{\pi} \\ 2, & \frac{1}{\pi} \le x \le 1. \end{cases}$$

So, the exact solution should be

$$u_e = \begin{cases} -\frac{1}{2}x^2 + \frac{5\pi^2+1}{2\pi(\pi+1)}x, & 0 \le x < \frac{1}{\pi} \\ -\frac{1}{4}x^2 + \frac{5\pi^2+1}{4\pi(\pi+1)}x + \frac{5\pi-1}{4\pi(\pi+1)}, & \frac{1}{\pi} \le x \le 1. \end{cases}$$

Iter	N_{elem}	$\|u-u_h\|_{L^\infty}$	$\|u-u_h\|_{L^2}$	$\|u-u_h\|_{H^1}$
1	2	5.652041×10^{-1}	4.506966×10^{-1}	9.637043×10^{-2}
2	4	5.652041×10^{-1}	4.626630×10^{-1}	4.818522×10^{-2}
3	8	5.652041×10^{-1}	4.656590×10^{-1}	2.409261×10^{-2}
4	16	5.652041×10^{-1}	4.664083×10^{-1}	1.204630×10^{-2}
5	32	5.652041×10^{-1}	4.665956×10^{-1}	6.023152×10^{-3}
6	48	5.652041×10^{-1}	4.666425×10^{-1}	4.116248×10^{-3}
7	96	5.652041×10^{-1}	4.666542×10^{-1}	2.058124×10^{-3}
8	160	5.652041×10^{-1}	4.666571×10^{-1}	1.739956×10^{-3}
9	192	5.652041×10^{-1}	4.666571×10^{-1}	1.029062×10^{-3}
10	192	5.652041×10^{-1}	4.666571×10^{-1}	1.029062×10^{-3}

Table 6: L^2 and H^1 Errors of Finite Element Method for Interface Problems .

Using linear regression, we can also see that the errors (Figure.8) in Table.6 obey

$$\|u-u_h\|_{H^1} \approx 0.1825 N_{elem}^{-0.9706}.$$

These linear regressions indicate that the adaptive finite element method for this problem can converge in the optimal rates, which is first order in H^1 norm.

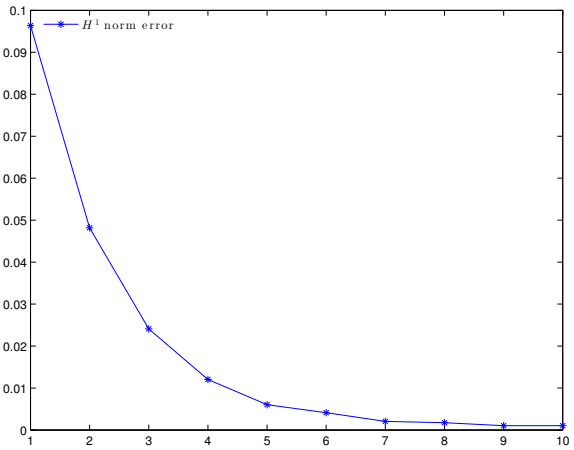

Figure 8: L^2 and H^1 norm errors for each iteration

Heat Equation

Problem 3

1. **Partition**

   I consider the following partition for finite element method:

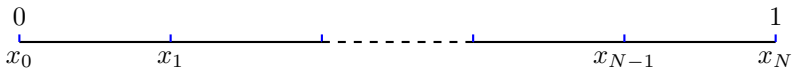

Figure 9: One dimension's uniform partition for finite element method

2. **The corresponding value of f and u_0**

   I choose the following corresponding value of f and u_0:

$$u_0 = \sin(3\pi x), f(t,x) = -2\sin(3\pi x)e^{-2t} + 9\pi^2 \sin(3\pi x)e^{-2t}.$$

3. **θ Method Scheme** The θ Method Discretization Scheme of this problem is as following

$$\frac{U_i^{k+1} - U_i^k}{\tau} - \theta\frac{U_{i-1}^k - 2U_i^k + U_{i+1}^k}{h^2} - (1-\theta)\frac{U_{i-1}^{k+1} - 2U_i^{k+1} + U_{i+1}^{k+1}}{h^2} = \theta f_i^k + (1-\theta)f_i^{k+1}. \tag{9}$$

   Let $\mu = \frac{\tau}{h^2}$, then the scheme (9) can be rewritten as

$$U_i^{k+1} - U_i^k - \theta\mu(U_{i-1}^k - 2U_i^k + U_{i+1}^k) - (1-\theta)\mu(U_{i-1}^{k+1} - 2U_i^{k+1} + U_{i+1}^{k+1}) = \theta\tau f_i^k + (1-\theta)\tau f_i^{k+1}.$$

   Combining of similar terms, we get

$$\begin{aligned} & -(1-\theta)\mu U_{i-1}^{k+1} + (2(1-\theta)\mu + 1)U_i^{k+1} - (1-\theta)\mu U_{i+1}^{k+1} \\ = \quad & \theta\mu U_{i-1}^k - (2\theta\mu - 1)U_i^k + \theta\mu U_{i+1}^k + \theta\tau f_i^k + (1-\theta)\tau f_i^{k+1}. \end{aligned}$$

   Since $U(0) = U(1) = 0$So, the θ-scheme can be written as the following matrix form

$$AU^{k+1} = BU^k + F,$$

   where

$$A = \begin{pmatrix} 2(1-\theta)\mu + 1 & -(1-\theta)\mu & & & & \\ -(1-\theta)\mu & 2(1-\theta)\mu + 1 & -(1-\theta)\mu & & & \\ & \ddots & \ddots & \ddots & & \\ & & -(1-\theta)\mu & 2(1-\theta)\mu + 1 & -(1-\theta)\mu & \\ & & & -(1-\theta)\mu & 2(1-\theta)\mu + 1 & \end{pmatrix},$$

$$B = \begin{pmatrix} -(2\theta\mu - 1) & \theta\mu & & & & \\ \theta\mu & -(2\theta\mu - 1) & \theta\mu & & & \\ & \ddots & \ddots & \ddots & & \\ & & \theta\mu & -(2\theta\mu - 1) & \theta\mu & \\ & & & \theta\mu & -(2\theta\mu - 1) & \end{pmatrix},$$

$$U^{k+1} = \begin{pmatrix} U^{k+1}(x_1) \\ U^{k+1}(x_2) \\ \vdots \\ U^{k+1}(x_{N-2}) \\ U^{k+1}(x_{N-1}) \end{pmatrix}, U^k = \begin{pmatrix} U^k(x_1) \\ U^k(x_2) \\ \vdots \\ U^k(x_{N-2}) \\ U^k(x_{N-1}) \end{pmatrix},$$

$$F = \theta\tau \begin{pmatrix} f^k(x_1) \\ \vdots \\ f^k(x_i) \\ \vdots \\ f^k(x_{N-1}) \end{pmatrix} + (1-\theta)\tau \begin{pmatrix} f^{k+1}(x_1) \\ \vdots \\ f^{k+1}(x_i) \\ \vdots \\ f^{k+1}(x_{N-1}) \end{pmatrix}.$$

4. **Numerical Results of Finite difference Method (θ Method) for Heat Equation**

(a) Numerical results for θ-Method for fixed $\tau = 1 \times 10^{-5}$

h	N_{nodes}	μ	$\|u-u_h\|_{L^\infty,\theta=0}$	$\|u-u_h\|_{L^\infty,\theta=1}$	$\|u-u_h\|_{L^\infty,\theta=\frac{1}{2}}$
1/4	5	0.00016	8.794539×10^{-2}	8.794522×10^{-2}	8.794531×10^{-2}
1/8	9	0.00064	1.723827×10^{-2}	1.723819×10^{-2}	1.723823×10^{-2}
1/16	17	0.00256	4.076556×10^{-3}	4.076490×10^{-3}	4.076523×10^{-3}
1/32	33	0.01024	1.005390×10^{-3}	1.005327×10^{-3}	1.005359×10^{-3}
1/64	65	0.04096	2.505219×10^{-4}	2.504594×10^{-4}	2.532024×10^{-4}
1/128	129	0.16384	6.260098×10^{-5}	6.253858×10^{-5}	6.256978×10^{-5}

Table 7: L^∞ norms for the θ-Method for fixed $\tau = 1 \times 10^{-5}$

h	N_{nodes}	μ	$\|u-u_h\|_{L^2,\theta=0}$	$\|u-u_h\|_{L^2,\theta=1}$	$\|u-u_h\|_{L^2,\theta=\frac{1}{2}}$
1/4	5	0.00016	6.218678×10^{-2}	6.218666×10^{-2}	6.218672×10^{-2}
1/8	9	0.00064	1.218929×10^{-2}	1.218924×10^{-2}	1.218927×10^{-2}
1/16	17	0.00256	2.882561×10^{-3}	2.882514×10^{-3}	2.882537×10^{-3}
1/32	33	0.01024	7.109183×10^{-4}	7.108736×10^{-4}	7.108959×10^{-4}
1/64	65	0.04096	1.771458×10^{-4}	1.771015×10^{-4}	1.771236×10^{-4}
1/128	129	0.16384	4.426558×10^{-5}	4.422145×10^{-5}	4.424352×10^{-5}

Table 8: L^2 norms for the θ-Method for fixed $\tau = 1 \times 10^{-5}$

h	N_{nodes}	μ	$\|u-u_h\|_{H^1,\theta=0}$	$\|u-u_h\|_{H^1,\theta=1}$	$\|u-u_h\|_{H^1,\theta=\frac{1}{2}}$
1/4	5	0.00016	1.838499×10^{-0}	1.838496×10^{-0}	1.838497×10^{-0}
1/8	9	0.00064	8.668172×10^{-1}	8.668132×10^{-1}	8.668152×10^{-1}
1/16	17	0.00256	4.284228×10^{-1}	4.284158×10^{-1}	4.284193×10^{-1}
1/32	33	0.01024	2.136338×10^{-1}	2.136204×10^{-1}	2.136271×10^{-1}
1/64	65	0.04096	1.067553×10^{-1}	1.067286×10^{-1}	1.067419×10^{-1}
1/128	129	0.16384	5.338867×10^{-2}	5.333545×10^{-2}	5.336206×10^{-2}

Table 9: H^1 norms for the θ-Method for fixed $\tau = 1 \times 10^{-5}$

From the Table(7)-(9), we can conclude that when $\mu < 0.5$, Implicit Euler method, Explicit Euler method and Crank-Nicolson method are convergent with optimal order in spacial, which are second order in L^∞, L^2 norm and first order in H^1 norm.

(b) Numerical results for θ-Method for $\tau = \frac{1}{\sqrt{2}}$

h	N_{nodes}	μ	$\|u-u_h\|_{L^\infty}$	$\|u-u_h\|_{L^2}$	$\|u-u_h\|_{H^1}$
1/4	5	8.00	9.334285×10^{-2}	6.600336×10^{-2}	1.951333×10^{0}
1/8	9	22.63	1.418498×10^{-1}	1.003029×10^{-1}	7.132843×10^{0}
1/16	17	64.00	5.067314×10^{-3}	3.583132×10^{-3}	5.325457×10^{-1}
1/32	33	181.02	3.744691×10^{-2}	2.647897×10^{-2}	7.957035×10^{0}
1/64	65	512	6.776843×10^{-4}	4.791952×10^{-4}	2.887826×10^{-1}
1/128	129	1228.15	8.093502×10^{-3}	5.722970×10^{-3}	6.902469×10^{0}
1/256	257	4096	2.192061×10^{-4}	1.550021×10^{-4}	3.739592×10^{-2}

Table 10: Error norms for the Implicit Euler method with $\tau = \sqrt{h}$

h	N_{nodes}	μ	$\|u-u_h\|_{L^\infty}$	$\|u-u_h\|_{L^2}$	$\|u-u_h\|_{H^1}$
1/4	5	8.00	4.341161×10^{2}	3.069664×10^{2}	9.075199×10^{3}
1/8	9	22.63	8.631363×10^{1}	6.103296×10^{1}	4.340236×10^{3}
1/16	17	64.00	4.466761×10^{3}	3.158477×10^{3}	4.694310×10^{5}
1/32	33	181.02	2.482730×10^{3}	1.755559×10^{3}	5.275526×10^{5}
1/64	65	512	5.556307×10^{10}	2.439517×10^{10}	1.962496×10^{14}
1/128	129	1228.15	4.383362×10^{25}	1.193837×10^{25}	3.823127×10^{29}
1/256	257	4096	3.530479×10^{51}	1.095038×10^{51}	1.420743×10^{56}

Table 11: Error norms for the Explicit Euler method with $\tau = \sqrt{h}$

h	N_{nodes}	μ	$\|u-u_h\|_{L^\infty}$	$\|u-u_h\|_{L^2}$	$\|u-u_h\|_{H^1}$
1/4	5	8.00	3.937504×10^{-1}	2.784236×10^{-1}	8.231355×10^{0}
1/8	9	22.63	4.372744×10^{-2}	3.091997×10^{-2}	2.198812×10^{0}
1/16	17	64.00	1.007102×10^{-2}	7.121285×10^{-3}	1.058406×10^{0}
1/32	33	181.02	3.858423×10^{-2}	2.728317×10^{-2}	8.198702×10^{0}
1/64	65	512	1.408511×10^{-4}	9.959676×10^{-5}	6.002108×10^{-2}
1/128	129	1228.15	7.776086×10^{-3}	5.498523×10^{-3}	6.631764×10^{0}
1/256	257	4096	1.158509×10^{-5}	8.191894×10^{-6}	1.976382×10^{-2}

Table 12: Error norms for the Crank-Nicolson method with $\tau = \sqrt{h}$

From the Table(10)-(12), we can conclude that Implicit Euler method and Crank-Nicolson method are unconditional stable, while when $\mu > \frac{1}{2}$ Explicit Euler method is not stable.

(c) Numerical results for θ-Method for $\tau = h$

h	N_{nodes}	μ	$\|u-u_h\|_{L^\infty}$	$\|u-u_h\|_{L^2}$	$\|u-u_h\|_{H^1}$
1/4	5	4	9.048357×10^{-2}	6.398155×10^{-2}	1.891560×10^{0}
1/8	9	8	1.777939×10^{-2}	1.257192×10^{-2}	8.940271×10^{-1}
1/16	17	16	4.292498×10^{-3}	3.035255×10^{-3}	4.511170×10^{-1}
1/32	33	32	1.106397×10^{-3}	7.823405×10^{-4}	2.350965×10^{-1}
1/64	65	64	2.999114×10^{-4}	2.120694×10^{-4}	1.278017×10^{-1}
1/128	129	128	8.707869×10^{-5}	6.157393×10^{-5}	7.426427×10^{-2}
1/256	257	256	2.785209×10^{-5}	1.969440×10^{-5}	4.751484×10^{-2}

Table 13: Error norms for the Implicit Euler method with $\tau = h$

h	N_{nodes}	μ	$\|u-u_h\|_{L^\infty}$	$\|u-u_h\|_{L^2}$	$\|u-u_h\|_{H^1}$
1/4	5	4	1.633634×10^{4}	1.155154×10^{4}	3.415113×10^{5}
1/8	9	8	4.782087×10^{6}	3.381446×10^{6}	2.404647×10^{8}
1/16	17	16	3.367080×10^{12}	2.023268×10^{12}	1.028718×10^{15}
1/32	33	32	1.762004×10^{51}	8.628878×10^{50}	1.756719×10^{54}
1/64	65	64	5.115840×10^{137}	2.577582×10^{137}	2.101478×10^{141}
1/128	129	128	4.972138×10^{-17}	∞	∞
1/256	257	256	4.972138×10^{-17}	∞	∞

Table 14: Error norms for the Explicit Euler method with $\tau = h$

h	N_{nodes}	μ	$\|u-u_h\|_{L^\infty}$	$\|u-u_h\|_{L^2}$	$\|u-u_h\|_{H^1}$
1/4	5	4	1.115040×10^{-1}	7.884526×10^{-2}	2.330993×10^{0}
1/8	9	8	1.245553×10^{-2}	8.807388×10^{-3}	6.263197×10^{-1}
1/16	17	16	4.072106×10^{-3}	2.879414×10^{-3}	4.279551×10^{-1}
1/32	33	32	1.004329×10^{-3}	7.101680×10^{-4}	2.134083×10^{-1}
1/64	65	64	2.502360×10^{-4}	1.769436×10^{-4}	1.066335×10^{-1}
1/128	129	128	6.250630×10^{-5}	4.419863×10^{-5}	5.330792×10^{-2}
1/256	257	256	1.562328×10^{-5}	1.104733×10^{-5}	2.665286×10^{-2}

Table 15: Error norms for the Crank-Nicolson method with $\tau = h$

From the Table(13)-(15), we can conclude that Implicit Euler method and Crank-Nicolson method are unconditional stable, while when $\mu > \frac{1}{2}$ Explicit Euler method is not stable.

(d) Numerical results for θ-Method for $\tau = h^2$

h	N_{nodes}	μ	$\|u-u_h\|_{L^\infty}$	$\|u-u_h\|_{L^2}$	$\|u-u_h\|_{H^1}$
1/4	5	1	8.849982×10^{-2}	6.257882×10^{-2}	1.850089×10^{0}
1/8	9	1	1.730081×10^{-2}	1.223352×10^{-2}	8.699621×10^{-1}
1/16	17	1	4.089480×10^{-3}	2.891699×10^{-3}	4.297810×10^{-1}
1/32	33	1	1.008450×10^{-3}	7.130822×10^{-4}	2.142840×10^{-1}
1/64	65	1	2.512547×10^{-4}	1.776639×10^{-4}	1.070675×10^{-1}
1/128	129	1	6.276023×10^{-5}	4.437819×10^{-5}	5.352449×10^{-2}
1/256	257	1	1.568672×10^{-5}	1.109219×10^{-5}	2.676109×10^{-2}

Table 16: Error norms for the Implicit Euler method with $\tau = h^2$

h	N_{nodes}	μ	$\|u-u_h\|_{L^\infty}$	$\|u-u_h\|_{L^2}$	$\|u-u_h\|_{H^1}$
1/4	5	1	8.603950×10^{5}	6.083912×10^{5}	1.798656×10^{7}
1/8	9	1	8.967110×10^{12}	6.340704×10^{12}	7.960153×10^{14}
1/16	17	1	3.903063×10^{104}	2.759883×10^{104}	1.406256×10^{107}
1/32	33	1	4.972138×10^{-17}	∞	∞
1/64	65	1	4.972138×10^{-17}	∞	∞
1/128	129	1	4.972138×10^{-17}	∞	∞
1/256	257	1	4.972138×10^{-17}	∞	∞

Table 17: Error norms for the Explicit Euler method with $\tau = h^2$

h	N_{nodes}	μ	$\|u-u_h\|_{L^\infty}$	$\|u-u_h\|_{L^2}$	$\|u-u_h\|_{H^1}$
1/4	5	1	8.793428×10^{-2}	6.217892×10^{-2}	1.838267×10^{0}
1/8	9	1	1.723790×10^{-2}	1.218904×10^{-2}	8.667990×10^{-1}
1/16	17	1	4.076506×10^{-3}	2.882525×10^{-3}	4.284175×10^{-1}
1/32	33	1	1.005358×10^{-3}	7.108952×10^{-4}	2.136269×10^{-1}
1/64	65	1	2.504906×10^{-4}	1.771236×10^{-4}	1.067419×10^{-1}
1/128	129	1	6.256978×10^{-5}	4.424351×10^{-5}	5.336206×10^{-2}
1/256	257	1	1.563914×10^{-5}	1.105854×10^{-5}	2.667992×10^{-2}

Table 18: Error norms for the Crank-Nicolson method with $\tau = h^2$

From the Table(16)-(18), we can conclude that Implicit Euler method and Crank-Nicolson method are unconditional stable, while when $\mu > \frac{1}{2}$ Explicit Euler method is not stable. Moreover, by using linear regression (Figure.10) for Implicit Euler method errors, we can see that the errors in Table.16 obey

$$\|u-u_h\|_{L^2} \approx 0.9435 h^{2.0580},$$
$$\|u-u_h\|_{H^1} \approx 7.2858 h^{1.0137}.$$

These linear regressions indicate that the finite element method for this problem can converge in the optimal rates, which are second order in L^2 norm and first order in H^1 norm.

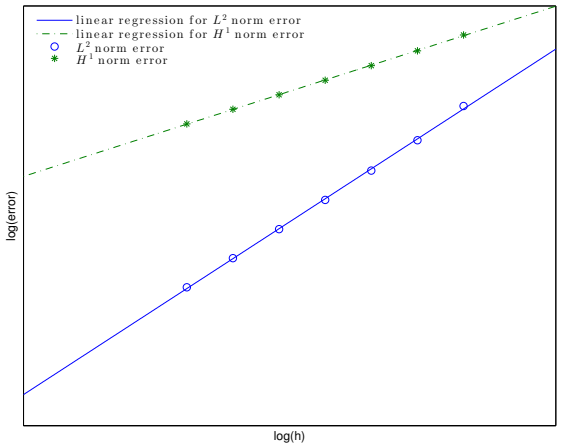

Figure 10: linear regression for L^2 and H^1 norm errors of Implicit Euler method with $\tau = h^2$

Similarly, by using linear regression (Figure.11) for Crank-Nicolson Method, we can also see that the errors in Table.18 obey

$$\|u-u_h\|_{L^2} \approx 0.9382 h^{2.0574},$$
$$\|u-u_h\|_{H^1} \approx 7.2445 h^{1.0131}.$$

These linear regressions indicate that the finite element method for this problem can converge in the optimal rates, which are second order in L^2 norm and first order in H^1 norm.

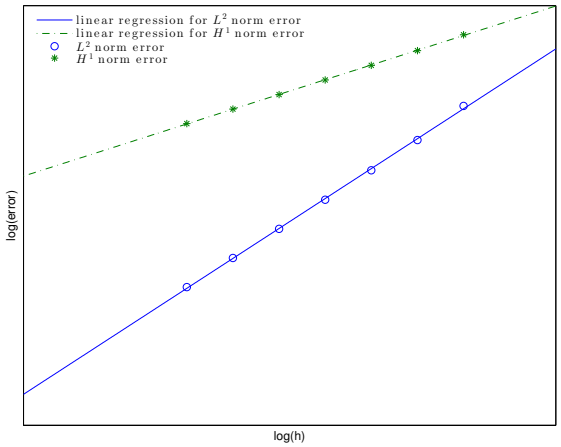

Figure 11: linear regression for L^2 and H^1 norm errors of Crank-Nicolson method with $\tau = h^2$