Prelim Notes for Numerical Analysis

Wenqiang Feng
wfeng1@utk.edu
Department of Mathematics,
University of Tennessee, Knoxville, TN, 37996

April 21, 2019

Abstract

This note is for my Numerical Analysis prelim exam in University of Tennessee at Knoxville. This note is intended to assist my prelim examination preparation. You can download and distribute it. Please be aware, however, that the note might contain typos as well as incorrect or inaccurate solutions . At here, I also would like to thank Liguo Wang for their help in some problems.

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*

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

11

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

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

1

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

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

2

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.*

3

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,*

4

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

5

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

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

6

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

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

7

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*

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

8

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 (1), 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 (1), then

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

Subtracting (3) form (2), 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 $A\mathbb{C}^{n\times n}$ is said to be unitary [a], if*

$$A^*A = I.$$

[a] A matrix $A\mathbb{R}^{n\times n}$ 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 (A) □

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 (A) □

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

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

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, $B = B^*$. □

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 1.17. *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 1.18. *$A \in \mathbb{C}^{n\times n}$ is normal if and only if every matrix unitarily equivalent to A is normal.*

Theorem 1.19. *$A \in \mathbb{C}^{n\times n}$ 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 1.20. *(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 1.21. *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

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

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 $(A^T y)^T x = y^T Ax = 0$ 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.*

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

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

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

Definition 1.10. *(Commonly used taylor series)*

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

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

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

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

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

Definition 1.11. *(Cauchy's Inequality)*

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

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

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

1

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

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

2

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

3

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

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

Discrete Version:

$$\left|\sum_{k=1}^n a_k b_k\right| \le \left(\sum_{k=1}^n |a_k|^p\right)^{1/p} \left(\sum_{k=1}^n |b_k|^q\right)^{1/q}. \tag{29}$$

General Version: Let $1<p_1,\cdots,p_n<\infty, \frac{1}{p_1}+\cdots+\frac{1}{p_n}=1$. Then if $u_k \in L^{p_k}(U)$, we have

$$\int_U |u_1 \cdots u_n| dx \le \Pi_{k=1}^n \|u_i\|_{L^{p_k}}(U). \tag{30}$$

4

Definition 1.16. *(Cauchy-Schwarz's Inequality) Let $1 \le p < \infty$ and $u,v \in L^p(U)$. Then*

$$|uv|^2 \le \|u\|_{L^2(U)} \|v\|_{L^2(U)}. \tag{31}$$

Discrete Version:

$$\left|\sum_{i=1}^n x_i y_i\right|^2 \le \sum_{i=1}^n |x_i|^2 \sum_{i=1}^n |y_i|^2. \tag{32}$$

5

Definition 1.17. *(Minkowski's Inequality)* *Let $1 \le p < \infty$ and $u, v \in L^p(U)$. Then*

$$\|u + v\|_{L^p(U)} = \|u\|_{L^p(U)} + \|v\|_{L^p(U)}. \tag{33}$$

Discrete Version:

$$\left(\sum_{k=1}^{n} |a_k + b_k|^p\right)^{1/p} \le \left(\sum_{k=1}^{n} |a_k|^p\right)^{1/p} + \left(\sum_{k=1}^{n} |b_k|^p\right)^{1/p}. \tag{34}$$

6

Definition 1.18. *(Gronwall's Inequality)*
Differential Version: Let $\eta(\cdot)$ be a nonnegative, absolutely continuous function on [0, T], which satisfies for a.e t the differential inequality

$$\eta'(t) \le \phi(t)\eta(t) + \psi(t), \tag{35}$$

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

$$\eta(t) \le e^{\int_0^t \phi(s)ds}\left[\eta(0) + \int_0^t \psi(s)ds\right], \forall 0 \le t \le T. \tag{36}$$

In particular, if

$$\eta' \le \phi\eta, \ on [0,T] \ and \ \eta(0) = 0, \tag{37}$$

$$\eta(t) = 0, \forall 0 \le t \le T. \tag{38}$$

Integral Version: Let $\xi(\cdot)$ be a nonnegative, summable function on [0, T], which satisfies for a.e t the integral inequality

$$\xi(t) \le C_1 \int_0^t \xi(s)ds + C_2, \tag{39}$$

where $C_1, C_2 \ge 0$. Then

$$\xi(t) \le C_2\left(1 + C_1 t e^{C_1 t}\right), \ \forall a.e. \ 0 \le t \le T. \tag{40}$$

In particular, if

$$\xi(t) \le C_1 \int_0^t \xi(s)ds, \ \forall a.e. \ 0 \le t \le T, \tag{41}$$

$$\xi(t) = 0, a.e. \tag{42}$$

Discrete Version: If

$$(1+\gamma)a_{n+1} \le a_n + \beta f_n, \ \alpha, \gamma \in \mathbb{R} \ \gamma \ne -1, \tag{43}$$

then,

$$a_{n+1} \le \frac{a_0}{(1+\gamma)^{n+1}} + \beta \sum_{k=0}^{n} \frac{f_k}{(1+\gamma)^{n+k-1}}. \tag{44}$$

7

1.3 Norms' Preliminaries

1.3.1 Vector Norms

Definition 1.19. *(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* : $\|x\| \geq 0, \quad (\|x\| = 0 \Leftrightarrow x = 0)$,
2. *homegenity* : $\|\alpha x\| = |\alpha| \|x\|$,
3. *triangle inequality* : $\|x + y\| \leq \|x\| + \|y\|, \quad \forall x, y \in \mathbb{R}^n$,

Definition 1.20. *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 1.3. *For all $x \in \mathbb{R}^n$,*

$$\|x\|_2 \leq \|x\|_1 \leq \sqrt{n} \|x\|_2, \tag{45}$$

$$\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n} \|x\|_\infty, \tag{46}$$

$$\frac{1}{\sqrt{n}} \|x\|_1 \leq \|x\|_2 \leq \sqrt{n} \|x\|_1, \tag{47}$$

$$\|x\|_\infty \leq \|x\|_1 \leq \sqrt{n} \|x\|_\infty. \tag{48}$$

Theorem 1.22. *(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*

$$\|Qx\|_2 = \|x\|_2, \quad \forall x \in \mathbb{R}^n \tag{49}$$

Proof.

$$\|Qx\|_2^2 = (Qx)^T (Qx) = x^T Q^T Q x = x^T x = \|x\|_2^2.$$

$\square$

1.3.2 Matrix Norms

Definition 1.21. *(Matrix Norms) A matrix norm is a function $\|\cdot\| : \mathbb{R}^{m\times n} \mapsto \mathbb{R}$ 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* : $\left\|x+y\right\| \leq \|x\| + \left\|y\right\|, \ \forall x, y \in \mathbb{R}^n$,

Definition 1.22. *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 1.4. *For all $A \in \mathbb{C}^{n\times n}$,*

$$\|A\|_2 \leq \|A\|_F \leq \sqrt{n}\,\|A\|_2\,, \tag{50}$$

$$\frac{1}{\sqrt{n}}\|A\|_2 \leq \|A\|_\infty \leq \sqrt{n}\,\|A\|_2\,, \tag{51}$$

$$\frac{1}{\sqrt{n}}\|A\|_\infty \leq \|A\|_2 \leq \sqrt{n}\,\|A\|_\infty\,, \tag{52}$$

$$\frac{1}{\sqrt{n}}\|A\|_1 \leq \|A\|_2 \leq \sqrt{n}\,\|A\|_1\,. \tag{53}$$

Corollary 1.5. *For all $A \in \mathbb{C}^{n\times n}$, then $\|A\|_2 \leq \sqrt{\|A\|_1\,\|A\|_\infty}$.*

Proof.

$$\|A\|_2^2 = \kappa(A) \leq \|A\|_1\left\|A^{-1}\right\|_1 = \|A\|_1\,\|A\|_\infty\,.$$

$\kappa(A)$: condition number ∎

Theorem 1.23. *(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*

$$\|QA\|_2 = \|A\|_2\,, \ \forall A \in \mathbb{R}^{n\times n}, \tag{54}$$

$$\|QA\|_F = \|A\|_F\,, \ \forall A \in \mathbb{R}^{n\times n} \tag{55}$$

Theorem 1.24. *Suppose that $A \in \mathbb{R}^{n\times n}$. If $\|A\| < 1$, then $(I-A)$ is nonsingular and*

$$(I-A)^{-1} = \sum_{k=0}^{\infty} A^k \tag{56}$$

with

$$\left\|(I-A)^{-1}\right\| \le \frac{1}{1-\|A\|}. \tag{57}$$

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

2 Direct Method

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

2.1.1 Singular Value Decomposition

2.1.2 QR Decomposition

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

2.2.1 LU Decomposition

2.2.2 Cholesky Decomposition

3 Iterative Method

3.1 General Iterative Scheme

3.2 Jacobi Method

3.3 Gauss-Seidel Method

3.4 Richardson Method

3.5 Successive Over Relaxation (SOR) Method

3.6 Minimal Correction Method

3.7 Steepest Descent Method

3.8 Conjugate Gradients Method

4 Eigenvalue Problems

5 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

$$|f(t,x) - f(t,y)| \leq \lambda |x - y|, \quad \lambda > 0. \tag{58}$$

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 5.1. *(Order of the Method) A time stepping scheme*

$$y_{n+1} = \Phi(h, y_0, y_1, \cdots, y_n) \tag{59}$$

is of order of $p \geq 1$, if

$$y_{n+1} - \Phi(h, y_0, y_1, \cdots, y_n) = \mathcal{O}(h^{p+1}). \tag{60}$$

Definition 5.2. *(Convergence of the Method) A time stepping scheme*

$$y_{n+1} = \Phi(h, y_0, y_1, \cdots, y_n) \tag{61}$$

is convergent , if

$$\lim_{h\to 0} \max_{n} \left\| y(t_n) - y_n \right\| = 0. \tag{62}$$

5.1 Euler's method

Definition 5.3. *(Forward Euler Method[a])*

$$y_{n+1} = y_n + hf(t_n, y_n), \qquad n = 0, 1, 2, \cdots. \tag{63}$$

[a]Forward Euler Method is explicit.

Theorem 5.1. *(Forward Euler Method is of order 1 [a]) Forward Euler Method*

$$y(t_{n+1}) = y(t_n) + hf(t_n, y(t_n)), \tag{64}$$

is of order 1 .

[a]You can also use multi-step theorem to derive it.

Proof. By the Taylor expansion,

$$y(t_{n+1}) = y(t_n) + hy'(t_n) + \mathcal{O}(h^2). \tag{65}$$

So,

$$\begin{aligned} y(t_{n+1}) - y(t_n) - hf(t_n, y(t_n)) &= y(t_n) + hy'(t_n) + \mathcal{O}(h^2) - y(t_n) - hf(t_n, y(t_n)) \\ &= y(t_n) + hy'(t_n) + \mathcal{O}(h^2) - y(t_n) - hy'(t_n) \\ &= \mathcal{O}(h^2). \end{aligned} \tag{66}$$

Therefore, Forward Euler Method (5.3) is order of 1 . □

Theorem 5.2. *(The convergence of Forward Euler Method) Forward Euler Method*

$$y(t_{n+1}) = y(t_n) + hf(t_n, y(t_n)), \tag{67}$$

is convergent.

Proof. From (66), we get

$$y(t_{n+1}) = y(t_n) + hf(t_n, y(t_n)) + \mathcal{O}(h^2), \tag{68}$$

Subtracting (68) from (63), we get

$$e_{n+1} = e_n + h[f(t_n, y_n) - f(t_n, y(t_n))] + ch^2. \tag{69}$$

Since f is lipschitz continuous w.r.t. the second variable, then

$$|f(t_n, y_n) - f(t_n, y(t_n))| \leq \lambda |y_n - y(t_n)|, \quad \lambda > 0. \tag{70}$$

Therefore,

$$\begin{aligned} \|e_{n+1}\| &\leq \|e_n\| + h\lambda \|e_n\| + ch^2 \\ &= (1 + h\lambda)\|e_n\| + ch^2. \end{aligned} \tag{71}$$

Claim:[?]

$$\|e_n\| \leq \frac{c}{\lambda} h[(1 + h\lambda)^n - 1], n = 0, 1, \cdots \tag{72}$$

Proof for Claim (72): 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:

$$\begin{aligned} \|e_{n+1}\| &\leq (1 + h\lambda)\|e_n\| + ch^2 & (73) \\ &\leq (1 + h\lambda)\frac{c}{\lambda} h[(1 + h\lambda)^n - 1] + ch^2 & (74) \\ &= \frac{c}{\lambda} h[(1 + h\lambda)^{n+1} - 1]. & (75) \end{aligned}$$

So, from the claim (72), we get $\|e_n\| \to 0$, when $h \to 0$. Therefore Forward Euler Method is convergent . ☐

Definition 5.4. *(Backward Euler Methods[a])*

$$y_{n+1} = y_n + hf(t_{n+1}, y_{n+1}), \qquad n = 0, 1, 2, \cdots. \tag{76}$$

[a] Backward Euler Method is implicit.

Theorem 5.3. *(backward Euler Method is of order 1 [a])* *Backward Euler Method*

$$y(t_{n+1}) = y(t_n) + hf(t_{n+1}, y(t_{n+1})), \tag{77}$$

is of order 1 .

[a] You can also use multi-step theorem to derive it.

Proof. By the Taylor expansion,

$$\begin{aligned} y(t_{n+1}) &= y(t_n) + hy'(t_n) + \mathcal{O}(h^2) & (78) \\ y'(t_{n+1}) &= y'(t_n) + \mathcal{O}(h). & (79) \end{aligned}$$

So,

$$\begin{aligned} & y(t_{n+1}) - y(t_n) - hf(t_{n+1}, y(t_{n+1})) \\ = \ & y(t_{n+1}) - y(t_n) + hy'(t_{n+1}) \\ = \ & y(t_n) + hy'(t_n) + \mathcal{O}(h^2) - y(t_n) - h[y'(t_n) + \mathcal{O}(h)] \\ = \ & \mathcal{O}(h^2). \end{aligned} \tag{80}$$

Therefore, Backward Euler Method (5.4) is order of 1 . □

Theorem 5.4. *(The convergence of Backward Euler Method)* *Backward Euler Method*

$$y(t_{n+1}) = y(t_n) + hf(t_{n+1}, y(t_{n+1})), \tag{81}$$

is convergent.

Proof. From (80), we get

$$y(t_{n+1}) = y(t_n) + hf(t_{n+1}, y(t_{n+1})) + \mathcal{O}(h^2), \tag{82}$$

Subtracting (82) from (76), we get

$$e_{n+1} = e_n + h[f(t_{n+1}, y_{n+1}) - f(t_{n+1}, y(t_{n+1}))] + ch^2. \tag{83}$$

Since f is lipschitz continuous w.r.t. the second variable, then

$$|f(t_{n+1}, y_{n+1}) - f(t_{n+1}, y(t_{n+1}))| \le \lambda |y_{n+1} - y(t_{n+1})|, \quad \lambda > 0. \tag{84}$$

Therefore,

$$\|e_{n+1}\| \le \|e_n\| + h\lambda \|e_{n+1}\| + ch^2. \tag{85}$$

So,

$$(1 - h\lambda)\|e_{n+1}\| \le \|e_n\| + ch^2. \tag{86}$$

So, by the Discrete Gronwall's Inequality , we have

$$\begin{aligned} \|e_{n+1}\| &\leq \frac{\|e_0\|}{(1-h\lambda)^{n+1}} + c\sum_{k=0}^{n} \frac{h^2}{(1-h\lambda)^{n+k-1}} \\ &= c\sum_{k=0}^{n} \frac{h^2}{(1-h\lambda)^{n+k-1}} \qquad (87) \\ &\leq ch^2(1+h\lambda)^{(nh)/h\lambda}(1-h\lambda \to 1+h\lambda) \\ &\leq che^T T. \end{aligned}$$

So, from the claim (87), we get $\|e_n\| \to 0$, when $h \to 0$. Therefore Forward Euler Method is convergent . $\square$

5.2 Trapezoidal Method

Definition 5.5. *(Trapezoidal Method[a])*

$$y_{n+1} = y_n + \frac{1}{2}h[f(t_n, y_n) + f(t_{n+1}, y_{n+1})], \qquad n = 0, 1, 2, \cdots. \tag{88}$$

[a]Trapezoidal Method Method is a combination of Forward Euler Method and Backward Euler Method.

Theorem 5.5. *(Trapezoidal Method is of order 2 [a]) Trapezoidal Method*

$$y(t_{n+1}) = y(t_n) + \frac{1}{2}h[f(t_n, y(t_n)) + f(t_{n+1}, y(t_{n+1}))], \tag{89}$$

is of order 2 .

[a]You can also use multi-step theorem to derive it.

Proof. By the Taylor expansion,

$$y(t_{n+1}) = y(t_n) + hy'(t_n) + \frac{1}{2!}h^2y''(t_n) + \mathcal{O}(h^3) \tag{90}$$

$$y'(t_{n+1}) = y'(t_n) + hy''(t_n) + \mathcal{O}(h^2). \tag{91}$$

So,

$$\begin{aligned} & y(t_{n+1}) - y(t_n) + \frac{1}{2}h[f(t_n, y(t_n)) + f(t_{n+1}, y(t_{n+1}))] \\ =\ & y(t_{n+1}) - y(t_n) + \frac{1}{2}h[y'(t_n) + y'(t_{n+1})] \\ =\ & y(t_n) + hy'(t_n) + \frac{1}{2!}h^2y''(t_n) + \mathcal{O}(h^3) - y(t_n) + \frac{1}{2}h[y'(t_n) + y'(t_n) + hy''(t_n) + \mathcal{O}(h^2)] \qquad (92) \\ =\ & \mathcal{O}(h^3). \end{aligned}$$

Therefore, Trapezoidal Method (5.5) is order of 2 . $\square$

Theorem 5.6. *(The convergence of Trapezoidal Method) Trapezoidal Method*

$$y(t_{n+1}) = y(t_n) + \frac{1}{2}h[f(t_n, y(t_n)) + f(t_{n+1}, y(t_{n+1}))], \tag{93}$$

is convergent.

Proof. From (92), we get

$$y(t_{n+1}) = y(t_n) + \frac{1}{2}h[f(t_n, y(t_n)) + f(t_{n+1}, y(t_{n+1}))] + \mathcal{O}(h^3), \tag{94}$$

Subtracting (94) from (88), we get

$$e_{n+1} = e_n + \frac{1}{2}h[f(t_n, y_n) - f(t_n, y(t_n)) + f(t_{n+1}, y_{n+1}) - f(t_{n+1}, y(t_{n+1}))] + ch^3. \tag{95}$$

Since f is lipschitz continuous w.r.t. the second variable, then

$$|f(t_n, y_n) - f(t_n, y(t_n))| \le \lambda |y_n - y(t_n)|, \quad \lambda > 0, \tag{96}$$

$$|f(t_{n+1}, y_{n+1}) - f(t_{n+1}, y(t_{n+1}))| \le \lambda |y_{n+1} - y(t_{n+1})|, \quad \lambda > 0. \tag{97}$$

Therefore,

$$\|e_{n+1}\| \quad \le \quad \|e_n\| + \frac{1}{2}h\lambda(\|e_n\| + \|e_{n+1}\|) + ch^3. \tag{98}$$

So,

$$(1 - \frac{1}{2}h\lambda)\|e_{n+1}\| \quad \le \quad (1 + \frac{1}{2}h\lambda)\|e_n\| + ch^3. \tag{99}$$

Claim:[?]

$$\|e_n\| \le \frac{c}{\lambda}h^2\left[\left(\frac{1+\frac{1}{2}h\lambda}{1-\frac{1}{2}h\lambda}\right)^n - 1\right], n = 0, 1, \cdots \tag{100}$$

Proof for Claim (100): The proof is by induction on n.

Then, we can make h small enough to such that $0 < h\lambda < 2$, then

$$\frac{1+\frac{1}{2}h\lambda}{1-\frac{1}{2}h\lambda} = 1 + \frac{1}{1-\frac{1}{2}h\lambda} \le \sum_{\ell=0}^{\infty}\frac{1}{\ell!}\left(\frac{h\lambda}{1-\frac{1}{2}h\lambda}\right)^{\ell} = exp\left(\frac{h\lambda}{1-\frac{1}{2}h\lambda}\right).$$

Therefore,

$$\|e_n\| \le \frac{c}{\lambda}h^2\left[\left(\frac{1+\frac{1}{2}h\lambda}{1-\frac{1}{2}h\lambda}\right)^n - 1\right] \le \frac{c}{\lambda}h^2\left(\frac{1+\frac{1}{2}h\lambda}{1-\frac{1}{2}h\lambda}\right)^n \le \frac{c}{\lambda}h^2 exp\left(\frac{nh\lambda}{1-\frac{1}{2}h\lambda}\right). \tag{101}$$

This bound of true for every negative integer n such that $nh < T$. Therefore,

$$\|e_n\| \le \frac{c}{\lambda}h^2 exp\left(\frac{nh\lambda}{1-\frac{1}{2}h\lambda}\right) \le \frac{c}{\lambda}h^2 exp\left(\frac{T\lambda}{1-\frac{1}{2}h\lambda}\right). \tag{102}$$

So, from the claim (102), we get $\|e_n\| \to 0$, when $h \to 0$. Therefore Forward Euler Method is convergent . □

5.3 Theta Method

Definition 5.6. *(Theta Method[a])*

$$y_{n+1} = y_n + \frac{1}{2}h[\theta f(t_n, y_n) + (1-\theta) f(t_{n+1}, y_{n+1})], \quad n = 0, 1, 2, \cdots. \tag{103}$$

[a]Theta Method is a general form of Forward Euler Method ($\theta = 1$), Backward Euler Method ($\theta = 0$) and Trapezoidal Method ($\theta = \frac{1}{2}$).

5.4 Midpoint Rule Method

Definition 5.7. *(Midpoint Rule Method)*

$$y_{n+1} = y_n + hf\left(t_n + \frac{1}{2}h, \frac{1}{2}(y_n + y_{n+1})\right). \tag{104}$$

Theorem 5.7. *(Midpoint Rule Method is of order 2) Midpoint Rule Method*

$$y(t_{n+1}) = y(t_n) + hf\left(t_n + \frac{1}{2}h, \frac{1}{2}(y(t_n) + y(t_{n+1}))\right). \tag{105}$$

is of order 2 .

Proof. By the Taylor expansion,

$$\begin{aligned} y(t_{n+1}) &= y(t_n) + hy'(t_n) + \frac{1}{2!}h^2 y''(t_n) + \mathcal{O}(h^3) \quad (106)\\ f(x_0 + \Delta x, y_0 + \Delta y) &= f(x_0, y_0) + \left(\Delta x \frac{\partial}{\partial x} + \Delta y \frac{\partial}{\partial y}\right) f(x_0, y_0) + \mathcal{O}(h^2). \quad (107)\end{aligned}$$

And chain rule

$$y'' = f'(t, \mathbf{y}) = \frac{\partial f(t, \mathbf{y})}{\partial t} + \frac{\partial f(t, \mathbf{y})}{\partial \mathbf{y}} f(t, \mathbf{y}). \tag{108}$$

So,

$$\begin{aligned} & y(t_{n+1}) - y(t_n) + hf\left(t_n + \frac{1}{2}h, \frac{1}{2}(y(t_n) + y(t_{n+1}))\right) \\ = \;& y(t_n) + hy'(t_n) + \frac{1}{2!}h^2 y''(t_n) + \mathcal{O}(h^3) - y(t_n) \\ - \;& h\left(f(t_n, y_n) + (t_n + \frac{1}{2}h - t_n)\frac{\partial f(t_n, y_n)}{\partial t} + (\frac{1}{2}(y(t_n) + y(t_{n+1})) - y_n)\frac{\partial f(t_n, y_n)}{\partial y} + \mathcal{O}(h^2)\right) \\ = \;& hy'(t_n) + \frac{1}{2!}h^2 y''(t_n) + \mathcal{O}(h^3) \\ - \;& \left(hf(t_n, y_n) + \frac{1}{2}h^2 \frac{\partial f(t_n, y_n)}{\partial t} + \frac{1}{2}h^2 \frac{\partial f(t_n, y_n)}{\partial y} + \mathcal{O}(h^3)\right) \end{aligned}$$

$$
\begin{aligned}
&= hy'(t_n)+\frac{1}{2!}h^2\left(\frac{\partial f(t_n,y_n)}{\partial t}+\frac{\partial f(t_n,y_n)}{\partial y}y'(t_n)\right)\\
&- \left(hy'(t_n)+\frac{1}{2}h^2\frac{\partial f(t_n,y_n)}{\partial t}+\frac{1}{2}h^2\frac{\partial f(t_n,y_n)}{\partial y}+\mathcal{O}(h^3)\right)\\
&= \mathcal{O}(h^3).
\end{aligned}
$$

Therefore, Midpoint Rule Method (5.5) is order of 2 . □

Theorem 5.8. *(The convergence of Midpoint Rule Method) Midpoint Rule Method*

$$y(t_{n+1}) = y(t_n)+hf\left(t_n+\frac{1}{2}h,\frac{1}{2}(y(t_n)+y(t_{n+1}))\right), \tag{109}$$

is convergent.

Proof. From (109), we get

$$y(t_{n+1}) = y(t_n)+hf\left(t_n+\frac{1}{2}h,\frac{1}{2}(y(t_n)+y(t_{n+1}))\right)+\mathcal{O}(h^3), \tag{110}$$

Subtracting (110) from (104), we get

$$e_{n+1}=e_n+h\left[f\left(t_n+\frac{1}{2}h,\frac{1}{2}(y(t_n)+y(t_{n+1}))\right)-f\left(t_n+\frac{1}{2}h,\frac{1}{2}(y(t_n)+y(t_{n+1}))\right)\right]+ch^3. \tag{111}$$

Since f is lipschitz continuous w.r.t. the second variable, then

$$
\begin{aligned}
&\left|f\left(t_n+\frac{1}{2}h,\frac{1}{2}(y(t_n)+y(t_{n+1}))\right)-f\left(t_n+\frac{1}{2}h,\frac{1}{2}(y(t_n)+y(t_{n+1}))\right)\right|\\
\leq\ & \frac{1}{2}\lambda|y_n-y(t_n)+y_{n+1}-y(t_{n+1})|,\quad \lambda>0.
\end{aligned} \tag{112}
$$

Therefore,

$$\|e_{n+1}\| \leq \|e_n\|+\frac{1}{2}h\lambda(\|e_n\|+\|e_{n+1}\|)+ch^3. \tag{113}$$

So,

$$(1-\frac{1}{2}h\lambda)\|e_{n+1}\| \leq (1+\frac{1}{2}h\lambda)\|e_n\|+ch^3. \tag{114}$$

Claim:[?]

$$\|e_n\|\leq\frac{c}{\lambda}h^2\left[\left(\frac{1+\frac{1}{2}h\lambda}{1-\frac{1}{2}h\lambda}\right)^n-1\right], n=0,1,\cdots \tag{115}$$

Proof for Claim (115): 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,

$$\|e_n\| \leq \frac{c}{\lambda}h^2\left[\left(\frac{1+\frac{1}{2}h\lambda}{1-\frac{1}{2}h\lambda}\right)^n - 1\right] \leq \frac{c}{\lambda}h^2\left(\frac{1+\frac{1}{2}h\lambda}{1-\frac{1}{2}h\lambda}\right)^n \leq \frac{c}{\lambda}h^2 exp\left(\frac{nh\lambda}{1-\frac{1}{2}h\lambda}\right). \tag{116}$$

This bound of true for every negative integer n such that $nh < T$. Therefore,

$$\|e_n\| \leq \frac{c}{\lambda}h^2 exp\left(\frac{nh\lambda}{1-\frac{1}{2}h\lambda}\right) \leq \frac{c}{\lambda}h^2 exp\left(\frac{T\lambda}{1-\frac{1}{2}h\lambda}\right). \tag{117}$$

So, from the claim (117), we get $\|e_n\| \to 0$, when $h \to 0$. Therefore Midpoint Rule Method is convergent . □

6 Multistep Methond

6.1 The Adams Method

Definition 6.1. *(s-step Adams-bashforth)*

$$y_{n+s} = y_{n+s-1} + h\sum_{m=0}^{s-1} b_m f(t_{n+m}, y_{n+m}), \tag{118}$$

where

$$\begin{aligned} b_m &= h^{-1}\int_{t_{n+s-1}}^{t_{n+s}} p_m(\tau)d\tau = h^{-1}\int_0^h p_m(t_{n+s-1}+\tau)d\tau \qquad n = 0,1,2,\cdots. \\ p_m(t) &= \Pi_{l=0,l\neq m}^{s-1}\frac{t-t_{n+l}}{t_{n+m}-t_{n+l}}, \quad \textit{Lagrange interpolation polynomials}. \end{aligned}$$

(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)

$$y_{n+3} = y_{n+2} + h\left[\frac{23}{12}f(t_{n+2}, y_{n+2}) - \frac{4}{3}f(t_{n+1}, y_{n+1}) + \frac{5}{12}f(t_n, y_n)\right].$$

281

6.2 The Order and Convergence of Multistep Methods

Definition 6.2. *(General s-step Method) The general s-step Method* [a] *can be written as*

$$\sum_{m=0}^{s} a_m \boldsymbol{y}_{n+m} = h \sum_{m=0}^{s} b_m \boldsymbol{f}(t_{n+m}, \boldsymbol{y}_{n+m}). \tag{119}$$

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.

282

Theorem 6.1. *(s-step method convergent order) The multistep method (119) is of order $p \geq 1$ if and only if there exists $c \neq 0$ s.t.*

$$\rho(w) - \sigma(w) \ln w^{a} = c(w-1)^{p+1} + \mathcal{O}(|w-1|^{p+2}), \quad w \to 1. \tag{120}$$

Where,

$$\rho(w) := \sum_{m=0}^{s} a_m w^m \quad \text{and} \quad \sigma(w) := \sum_{m=0}^{s} b_m w^m. \tag{121}$$

[a] Let $w = \xi + 1$, then $\ln(1+\xi) = \sum_{n=0}^{\infty} (-1)^n \frac{\xi^{n+1}}{n+1} = \xi - \frac{\xi^2}{2} + \frac{\xi^3}{3} - \frac{\xi^4}{4} + \cdots + (-1)^n \frac{\xi^{n+1}}{n+1} + \cdots, \xi \in (-1, 1)$.

283

Theorem 6.2. *(s-step method convergent order) The multistep method (119) 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,

$$\rho(w) := \sum_{m=0}^{s} a_m w^m \quad \text{and} \quad \sigma(w) := \sum_{m=0}^{s} b_m w^m. \tag{122}$$

284

Lemma 6.1. *(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.*

285

Theorem 6.3. *(The Dahlquist equivalence theorem) The multistep method (119) is convergent if and only if*

1. *consistency: multistep method (119)is order of $p \geq 1$,*
2. *stability: the polynomial $\rho(w)$ satisfies the root condition .*

286

6.3 Method of A-stable verification for Multistep Methods

Theorem 6.4. *Explicit Multistep Methods can not be A-stable.*

Theorem 6.5. *(Dahlquist second barrier) The highest oder of an A-stable multistep method is 2 .*

7 Runge-Kutta Methods

7.1 Quadrature Formulas

Definition 7.1. *(The Quadrature) The Quadrature is the procedure of replacing an integral with a finite sum.*

Definition 7.2. *(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

$$\int_a^b f(\tau)w(\tau)d\tau \approx \sum_j^n b_j f(c_j). \tag{123}$$

Remark 7.1. *The quadrature formula (123) is order of p if it is exact for every $f \in \mathbb{P}_{p-1}$.*

7.2 Explicit Runge-Kutta Formulas

7.3 Implicit Runge-Kutta Formulas

7.4 Collocation Runge-Kutta Formulas

7.5 Method of A-stable verification for Runge-Kutta Method

Theorem 7.1. *Explicit Runge-Kutta Methods can not be A-stable.*

7.6 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 (123) 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

$$\sum_{j=1}^{p} b_j c_j^m = \int_a^b \tau^m w(\tau)d\tau, \quad m=0,1,\cdots,p-1. \tag{124}$$

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

$$y(t_{n+1})-y(t_{n-1}) = \int_{t_{n-1}}^{t_{n+1}} f(s;y(s))ds. \tag{125}$$

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,

$$y(t_{n+1}) - y(t_{n-1}) = \int_{t_{n-1}}^{t_{n+1}} f(s;y(s))ds. \tag{126}$$

Then,by Simpson's quadrature rule, we have

$$\begin{aligned}
& y(t_{n+1}) - y(t_{n-1}) && (127)\\
= & \int_{t_{n-1}}^{t_{n+1}} f(s;y(s))ds. && (128)\\
= & \frac{t_{n+1}-t_{n-1}}{6}\left[f(t_{n-1};y(t_{n-1})) + 4f\left(\frac{t_{n-1}+t_{n+1}}{2};y\left(\frac{t_{n-1}+t_{n+1}}{2}\right)\right) + f(t_{n+1};y(t_{n+1}))\right] && (129)\\
= & \frac{h}{3}\left[f(t_{n-1};y(t_{n-1})) + 4f\left(t_n;y\left(t_n\right)\right) + f(t_{n+1};y(t_{n+1}))\right]. && (130)
\end{aligned}$$

Therefore, the 3-step method deriving from Simpson's rule is

$$y(t_{n+1}) = y(t_{n-1}) + \frac{h}{3}\left[f(t_{n-1};y(t_{n-1})) + 4f\left(t_n;y\left(t_n\right)\right) + f(t_{n+1};y(t_{n+1}))\right]. \tag{131}$$

Or

$$y(t_{n+2}) - y(t_n) = \frac{h}{3}\left[f(t_n;y(t_n)) + 4f\left(t_{n+1};y\left(t_{n+1}\right)\right) + f(t_{n+2};y(t_{n+2}))\right]. \tag{132}$$

2. The order For our this problem

$$\rho(w) := \sum_{m=0}^{s} a_m w^m = -1 + w^2 \quad \text{and} \quad \sigma(w) := \sum_{m=0}^{s} b_m w^m = \frac{1}{3} + \frac{4}{3}w + \frac{1}{3}w^2. \tag{133}$$

By making the substitution with $\xi = w - 1$ i.e. $w = \xi + 1$, then

$$\rho(w) := \sum_{m=0}^{s} a_m w^m = \xi^2 + 2\xi \quad \text{and} \quad \sigma(w) := \sum_{m=0}^{s} b_m w^m = \frac{1}{3}\xi^2 + 2\xi + 2. \tag{134}$$

So,

$$\begin{aligned}
\rho(w) - \sigma(w)ln(w) &= \xi^2 + 2\xi - (2 + 2\xi + \frac{1}{3}\xi^2)(\xi - \frac{\xi^2}{2} + \frac{\xi^3}{3}\cdots)\\
&= \begin{array}{lllll}
+2\xi & +\xi^2 & & & \\
-2\xi & +\xi^2 & -\frac{2}{3}\xi^3 & & \\
 & -2\xi^2 & +\xi^3 & -\frac{2}{3}\xi^4 & \\
 & & -\frac{1}{3}\xi^3 & +\frac{1}{6}\xi^4 & -\frac{1}{9}\xi^5
\end{array}\\
&= -\frac{1}{2}\xi^4 + \mathcal{O}(\xi^5).
\end{aligned}$$

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,

$$\rho(w) := \sum_{m=0}^{s} a_m w^m = -1 + w^2 = (w-1)(w+1). \tag{135}$$

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. ◀

Appendices

A Lecture notes

appendix1

B Trigonometric formula tables

C Trigonometric tables